Classification of Cured Tobacco Leaves by Colour and Plant Position by means of Computer Processing of Digital Images

George Metcalf Tattersfield

A dissertation submitted to the Department of Electrical Engineering,
University of Cape Town, in fulfilment of the requirements
for the degree of Master of Science in Engineering.

Cape Town, September 1999

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in Engineering in the University of Cape Town. It has not been submitted before for any degree or examination in any other university.

Signature of Author.... Signed by candidate .....

Cape Town
September 21, 1999

Abstract

This dissertation investigates the machine vision grading of flue-cured Virginia tobacco by means of digital processing of tobacco leaf images. With reference to international grading standards and to modern image processing techniques, two classifiers are designed. The colour classifier uses seven features extracted from each leaf image to grade the leaf into one of five official colour classes. It does this with an expected correct classification rate of 93.5%. The plant position classifier identifies the position on the stalk from which a leaf was reaped, using ten size and shape features to classify the leaf into one of six plant position categories. It has a correct classification rate of 70%. Average colours for each colour class and archetypal shapes for each plant position category are derived from the digital leaf data. These should be of value to tobacco graders as objective representations of typical leaves within each class.

This work is dedicated to my Uncle and Aunt

Rex and Sheila Tattersfield

who first suggested the project to me some years ago,

and who have encouraged me towards its completion ever since.

Acknowledgements

I am very grateful to my colleagues in the Department of Electrical Engineering for having afforded me the time from my teaching duties and the travel funds that made the researching of this project possible.

I am also much indebted to Mr L.T.V. Cousins, Director of the Tobacco Research Board in Zimbabwe, and to his staff, for extending to me the use of their excellent grading facilities and library at Kutsaga. I was assisted by several of the staff at the TRB, but would like to single out Rufus Pswarayi, a Farm Technical Assistant, and Mackson Nyambure, a Grader, who each gave me many hours of enthusiastic help during the photographing of the leaf samples.

The leaf material for this project was made available through the kind offices of Mr Stanley Mutepfa, General Manager of the Tobacco Industry Marketing Board in Harare. I am also particularly grateful for the advice and assistance that I received in collecting and grading the samples from Mr Enoch Bile, Chief Classifier (Flue Cured Tobacco) at the TIMB.

Within the University of Cape Town, I would like to thank my supervisor, Professor Gerhard de Jager, for the opportunity to work here and research in his field of image processing. I am also grateful to Assoc. Prof. Trevor Sewell, Director of the Electron Microscope Unit, for permission to use the slide scanner on several occasions.

During the time that I have spent working on this project, I have been fortunate to have had many advisors, colleagues and friends whose wisdom, ideas and support have buoyed my efforts, and for whose contributions to my life and work I am ever grateful. Difficult though it is to mention a particular few, I would wish to record my especial gratitude to Norman Ballard, Mark Cammidge, Keith Forbes, Hylton Gifford and Ian Greager, as well as to my family in the UK, Zimbabwe and Cape Town.

Contents

List of Figures

List of Tables

List of Symbols

a	—	An element within the set $\mathcal{A}$
$\mathbf{a}$	—	Unit vector in the direction of an axis in the feature space
$\mathbf{a}^T$	—	The transpose of the vector $\mathbf{a}$
b	—	An element within the set $\mathcal{B}$
d_c	—	Distance of a feature vector from the centroid of the cluster to which it is assigned
d_r	—	Distance from a feature vector to the centroid of class r
d^2	—	Squared Mahalanobis distance
$d_1(\mathbf{x})$, $d_2(\mathbf{x})$	—	Decision functions for classes 1 and 2
$\mathbf{e}_1$	—	The eigenvector corresponding to the eigenvalue λ_1
$f(\theta)$	—	The boundary function of an object in terms of the angle at the object's centroid
$f(t)$	—	Boundary function of an object in terms of a time-like variable
$f(x, y)$	—	An image with pixel values expressed as a function of the spatial co-ordinates
$f'(x, y)$	—	A thresholded image
$f'(x, y)$	—	A histogram equalised image
g	—	Index variable for classes in the feature space
g_1, g_2	—	Graylevels to which a thresholded image is set
g_r	—	A discrete array of sampled values of an object's boundary function
$g(x, y)$	—	The output image from a transformation such as translation, rotation or scaling
$g(x_c, y_c)$	—	The output value from a filter mask operation
$g_{(x,y)}$	—	Grayscale value at the point (x,y) in a grayscale image
$g'_{(x,y)}$	—	Grayscale value at the point (x,y) in a grayscale image after histogram equalisation
i	—	Index variable for summation of pixels within objects
i	—	Index variable for classes in feature space
i'	—	Index variable used like i, but primed to distinguish it from i
j	—	Index variable for summation of pixels within objects
j	—	Index variable for classes in feature space
j	—	$\sqrt{-1}$
k	—	Index variable for clusters in feature space

x, y, z	—	The trichromatic coefficients
y_{max}	—	The maximum vertical co-ordinate within an object
y_{min}	—	The minimum vertical co-ordinate within an object
$\bar{y}$	—	The mean value of the variable y
$y(t)$	—	The imaginary, y-component of the time-domain boundary function $f(t)$
$y(\theta)$	—	The imaginary, y-component of the boundary function $f(\theta)$
A	—	Extra factor grading symbol for spotted tobacco
A	—	Grading symbol for tobacco strips
A	—	A pixel point within an image
$\mathcal{A}$	—	A set of elements comprising an object within an image
det(**A**) or $\lvert\mathbf{A}\rvert$	—	The determinant of the matrix **A**
$\overline{AB}$	—	The distance in pixels from A to B
B	—	Grading symbol for tobacco scrap
B	—	A pixel point within an image
B	—	The blue component in the RGB model
$\bar{B}$	—	The mean value of the B colour component within an object
$\mathcal{B}$	—	A set of elements comprising an object within an image
$\mathcal{B}_x$	—	Translated version of $\mathcal{B}$, shifted by the vector offset x
$\widehat{BAC}$	—	The angle subtended at point A by lines AB and AC
$B(\lambda), G(\lambda), R(\lambda)$	—	Cone sensitivities of the human eye as functions of wavelength
C	—	Grading plant position symbol for cutters
C	—	A covariance matrix
C_1, C_2, C_3, C_4	—	The elements of a 2×2 covariance matrix
$\mathbf{C}_k$	—	The covariance matrix for cluster k
C_{jk}	—	Element in row j and column k of a covariance matrix
D	—	Extra factor grading symbol for "harsh-natured" tobacco
E	—	Grading colour symbol for pale lemon tobacco
F	—	Extra factor grading symbol for ripe tobacco
F	—	Statistic for determining the significance of a value of λ
$\mathcal{F}$	—	The Fourier transform operator
$\mathcal{F}^{-1}$	—	The inverse Fourier transform operator
$F(\omega)$	—	The Fourier transform of an object's boundary function
$F_s(\omega)$	—	The Fourier transform of a sampled boundary function
G	—	Extra factor grading symbol for green tobacco
G	—	The green component in the RGB model
$\bar{G}$	—	The mean value of the G colour component within an object
G_{md}	—	The modal value of the green histogram of the pixels within an object
G_{std}	—	The standard deviation of the green band histogram of an object ($=\sigma_G$)

G_{skew} — The skewness of the green band histogram of an object ($=\mu_{3G}/\sigma_G^3$)

G_{kurt} — The kurtosis of the green band histogram of an object ($=\mu_{4G}/\sigma_G^4$)

G_i — Proportion of pixels of component value i in an image

G_u — The discrete Fourier transform of a sampled boundary function array

G_0 — The dc Fourier descriptor

$G_0, G_{\pm1}, G_{\pm2}\ldots$ — The Fourier descriptors of an object

$G'_{\pm2}, G'_{\pm3}\ldots$ — Normalised Fourier descriptors

$G(i)$ — A function of i used in simplifying a probability calculation

H — Grading plant position symbol for smoking leaf

$\mathbb{I}^2$ — Two-dimensional space for set-theoretic treatment of morphological operations

$\mathbf{I}$ — An identity matrix

I — The intensity component in the HSI model

$\bar{I}$ — The mean value of intensity, I, within an object

I_{md} — The modal value of the intensity histogram of the pixels within an object

I_{std} — The standard deviation of the intensity band histogram of an object ($=\sigma_I$)

I_{skew} — The skewness of the intensity band histogram of an object ($=\mu_{3I}/\sigma_I^3$)

I_{kurt} — The kurtosis of the intensity band histogram of an object ($=\mu_{4I}/\sigma_I^4$)

K — Extra factor grading symbol for immature tobacco

K — An intensity threshold

K — A constant used in simplifying a probability calculation

L — The total luminance perceived by the eye

L — Grading colour symbol for lemon tobacco

L — Grading plant position symbol for leaf tobacco

L — The general form of a discriminator based on within-groups and overall data variability

$L(\lambda)$ — Colour distribution of the light radiating or reflecting from an object

N — The total number of pixels within an object

N — The total number of objects within all classes

O — Grading colour symbol for orange tobacco

O — A set of pixels comprising an object to be translated, rotated or scaled

P — Grading plant position symbol for primings

$P(\omega_i)$ — *A priori* probability that an object chosen at random belongs in class i

$P(\omega_i|\mathbf{x})$ — *A posteriori* conditional probability that $\mathbf{x}$ belongs in class i

Q — Extra factor grading symbol for scorched tobacco

Q — The number of points in a cluster

R — Grading colour symbol for light mahogany tobacco

R — Subscript indicating a probability that is based on a "real world" distribution

R — The red component in the RGB model

$\bar{R}$ — The mean value of the R colour component within an object

R_{md} — The modal value of the red histogram of the pixels within an object

R_{std} — The standard deviation of the red band histogram of an object ($=\sigma_R$)

R_{skew} — The skewness of the red band histogram of an object ($=\mu_{3R}/\sigma_R^3$)

R_{kurt} — The kurtosis of the red band histogram of an object ($=\mu_{4R}/\sigma_R^4$)

$\text{Rect}\left(\frac{\omega}{\Omega}\right)$ — Windowing function in the frequency domain

S — Grading colour symbol for mahogany tobacco

S_x — Scaling factor in the x-direction

S_y — Scaling factor in the y-direction

T — Grading plant position symbol for tips

T — The time period taken to traverse an object's boundary

T — Subscript indicating a probability that is based on *a priori* knowledge of the training set

$\mathbf{T}$ — The total sums of squares and cross-products matrix

V — Extra factor grading symbol for temporarily greenish tobacco

$\mathbf{W}$ — The within-groups sums of squares and cross-products matrix

X — A variable

X — Grading plant position symbol for lugs

'X' — A feature vector requiring classification

X, Y, Z — The tristimulus values

Y — A variable

Y — Extra factor grading symbol for tobacco with guineafowl spot

$\delta_T(t)$ — The dirac delta train used for function sampling

θ — Rotation angle

θ — The angle from the centroid of an object to one of its boundary points

Λ — Wilks' lambda

λ — The partial lambda, denoting discriminatory value of an individual feature

λ — Wavelength

λ — The variable in the characteristic polynomial of a system of equations

λ_1 — The larger eigenvalue of a 2×2 system

μ_r — The r^{th} moment about the mean

μ_2 — The second moment about the mean : variance

μ_3 — The third moment about the mean : skewness

$\mu_{3R}, \mu_{3G}, \mu_{3I}$ — The skewness of an object's red, green and intensity histograms

μ_4 — The fourth moment about the mean : kurtosis

$\mu_{4R}, \mu_{4G}, \mu_{4I}$ — The kurtosis of an object's red, green and intensity histograms

μ'_r — The r^{th} moment about the origin

σ_B^2 — The variance of the blue histogram values of the pixels within an object

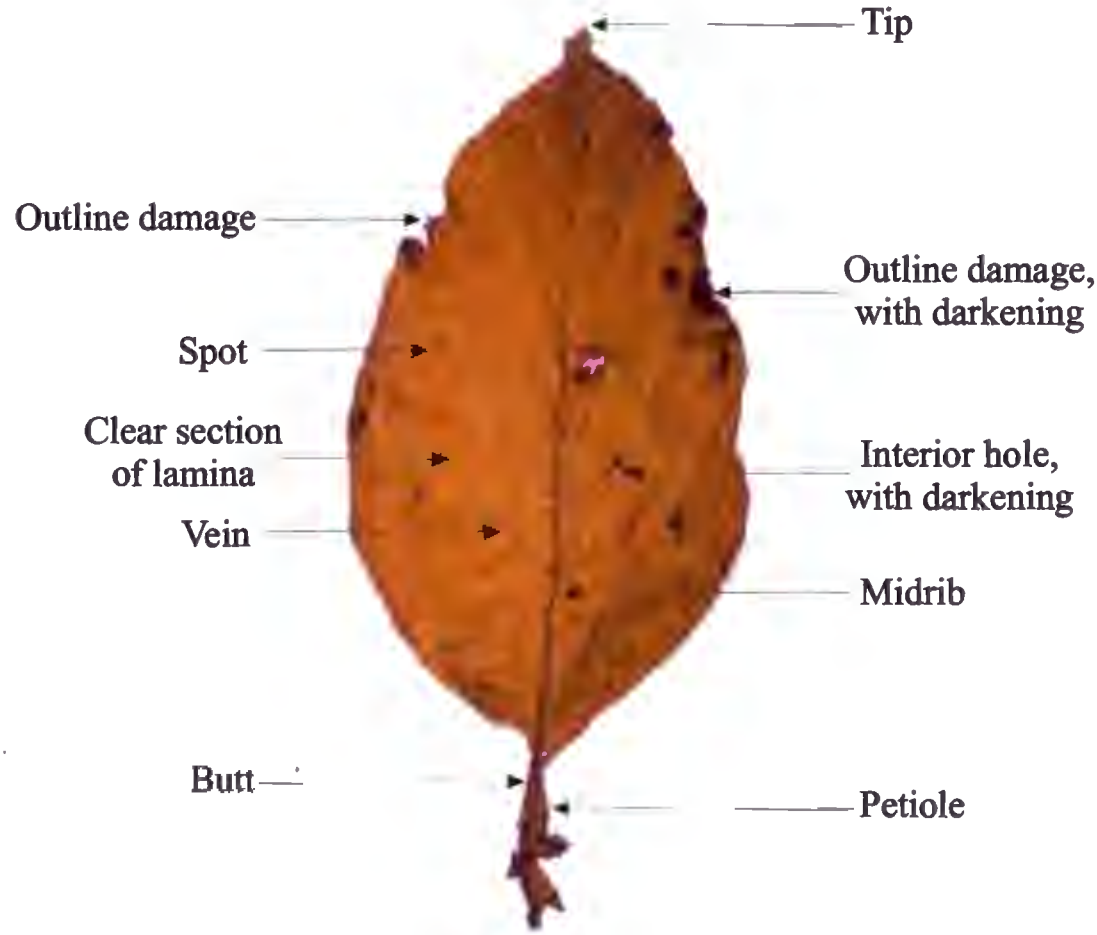
σ_G^2 — The variance of the green histogram values of the pixels within an object

σ_I^2 — The variance of the intensity histogram values of the pixels within an object

σ_R^2	—	The variance of the red histogram values of the pixels within an object
σ_X^2	—	Variance of the variable X
σ_{XY}^2	—	The covariance of the variables X and Y
σ_Y^2	—	Variance of the variable Y
$\phi(x)$	—	Probability density function of the variable x
ω	—	The frequency variable in Fourier analysis
$\omega_1, \omega_2, \ldots \omega_m$	—	The available classes in a classifier
Ω	—	A constant value of frequency used in frequency-domain windowing
$\odot$	—	The centroid of a cluster on a graph
$\oplus$	—	The dilation operator
$\ominus$	—	The erosion operator
$\circ$	—	The opening operator (erosion then dilation)
$\bullet$	—	The closing operator (dilation then erosion)
$+$	—	Vector addition of offsets of non-zero elements of a set, used in morphological operations
$\cap$	—	Intersection of sets
$\cup$	—	Union of sets
$\in$	—	Denotes that an element belongs to a set
$\emptyset$	—	The null set

Nomenclature

Technical terms in tobacco grading or image processing are *italicised* and fully defined where they first appear in the text. To assist the reader, the names of various parts of a tobacco leaf are given here as a reference in terms of the leaf image below.



Ripe lemon cutter, graded C3LF

Chapter 1

Introduction

1.1 Historical context

Upon their arrival in the Americas in October 1492, Spanish explorers soon noted that the Indian inhabitants would burn certain leaves and inhale the smoke through hollow reeds for recreation. There is still disagreement [53] as to whether the word *tobacco* was derived, as Oveido claimed in 1535 [52], from the Y-shaped tubes or pipes which the Indians used for this purpose, or whether the word referred to the tubular rolls of leaves being burned [25]. Whichever the case, the practice of smoking tobacco leaves was adopted enthusiastically by the Spaniards, who took both the habit and good supplies of the leaf with them when they sailed for home. Rodrigo de Jerez, one of Christopher Columbus' lieutenants, is said to have been the first man to have smoked tobacco in Europe, an action which was denounced as "devilish" by the Inquisition and which swiftly led to his imprisonment in his home town of Ayamonte [15, 23].

Despite such formidable opposition, the smoking of tobacco flourished in Spain in the early 16th century, and spread during the next seventy-five years to other countries of Europe, particularly to Belgium, France and England (where the first seeds were imported in 1565) [7]. A notable early enthusiast was Jean (or Jacques) Nicot (c. 1530-1600), who encountered tobacco during his term as the French ambassador to the royal court at Lisbon and who first introduced tobacco smoking to France upon his return home in 1560 [62]. Nicot encouraged the use of tobacco for medicinal purposes, and such was his success that his name rapidly became synonymous with

the leaves, then with the action of smoking them and, much more recently, with the active chemical alkaloid ingredient, nicotine ($C_{10}H_{14}N_2$)[13]. The taxonomic plant genus *Nicotiana* includes two species that are particularly high in nicotine compared with wild tobacco [34]: *N. rustica* which originated in Mexico and was smoked in England until 1616; and *N. tabacum* which came from Brazil, was smoked by the Spanish and has been the tobacco of choice for most smokers since it was planted in Virginia in the 17th century. More obscurely, the adjectives *nicotian* and *nicotiant* have been applied to the leaves and to those that smoke them respectively, whilst the term *nicotism* is used, evocatively, for addictive indulgence in tobacco usage [53].

In the days of Sir Walter Raleigh, its great English proponent at the dawn of the 17th century, tobacco leaf sold for its weight in silver. In November 1593, Arthur Throckmorton bought $3\frac{1}{2}$ ounces from Raleigh (who was his brother-in-law) for eleven shillings and sixpence — a huge sum to pay, even for a rich man's luxury [57]. The King, James I, was implacably opposed to smoking and famously derided it in 1604 as

> A custome lothsome to the eye, hatefull to the Nose, harmefull to the braine, dangerous to the Lungs, and the blacke stinking fume thereof, neerest resembling the horrible Stigian smoke of the pit that is bottomlesse [42].

Nevertheless, by 1620 prices were beginning to fall and the weight of leaf imported annually from Virginia to England had reached 40 000 lbs (or about 18 t) [23, 46]. Furthermore, the King was already earning substantial duties from the trade, a fact which introduced an irony in the attitude of government towards smoking that still persists today. At the Restoration in 1660, King Charles II earned £400,000 annually from tobacco revenue [96], the importance of which is evident in comparison with his annual income from Parliament of £1.2 million and his debt of £925,000 at the time [99].

Despite stern opposition from authorities, including the threats at various times of excommunication from the Roman Catholic Church (by the Papal Bulls of 1624, 1642 and 1650), decapitation in China [7], transportation or death in orthodox Russia [35] and torture by means of a pipe inserted through the nose in Turkey [23], the demand for tobacco continued to grow rapidly. This was fuelled in part by a widely-held belief in its medicinal power: in a curious inversion of today's standards, for example, it

was made compulsory in 1665 for all the boys at Eton to smoke each morning as a means of warding off the Great Plague [5]. Other dubiously-advantageous uses of tobacco included the tobacco-smoke enema-syringe, which persisted through the 17th and 18th centuries as a major medicinal technique for resuscitating people in a state of suspended animation, or apparently drowned [10].

Patterns of tobacco usage since the 17th century have been determined far more by styles and fashions than by any mistaken faith in its curative properties. Snuff-taking was fashionable from around 1680 and throughout the 1700s, while in 1804 cigars made their first appearance in the United Kingdom from Spain. The smaller and cheaper cigarette followed in about 1842-3, beginning as an exclusive craze, "quite *la grande mode* of late with certain French ladies" [14], but rapidly becoming available to the masses so that soon it was reported that even "the beggars in the streets have paper cigars (called cigarettes) in their mouths" [56]. At the opposite end of the social scale, Queen Victoria was strongly opposed to the use of tobacco in all its forms [5, 15]. Nevertheless, the European wars of the late 19th and early 20th centuries catalysed the spread of cigarette smoking, with fashions being adopted by army officers and men through exposure to cultures that were hitherto unknown to them. So, for instance, the *Evening News* of 10th October, 1914 charges perhaps rather simplistically that "[o]ur officers ... brought the habit back with them from the Crimea, where they learned it from the Russians" [26].

By the mid-20th century, cigarettes had become universally available, and demand for them was augmented through their frequent use by the film stars of cinema and also through more overt forms of product advertisement. It has been estimated that in 1957 the *average* American citizen above 15 years of age smoked 3440 cigarettes per year, with corresponding figures of 2720 in Canada, 2630 in the United Kingdom, 2380 in Ireland and 1510 in Rhodesia [8]. Current annual world cigarette production stands at 5700 billion pieces, or just under 1000 cigarettes for every human being alive; and, despite greatly increased awareness of the dangers of smoking, consumption in the developing world continues to rise [102]. Adult annual cigarette consumption in Greece and Cyprus, the highest in the world, is still well over 3000. [24, 91].

Even in the face of the emotionally negative response which tobacco production inspires in many people today, the tobacco industry remains a vast worldwide enterprise which seeks to satisfy the ever-shifting fashions and demands of huge numbers of consumers. Technology has been used in the service of this industry ever since the

galleons, marvels of their time, set out across the ocean to the Americas. Very efficient mass production techniques and high-precision chemical analysis are just two of the technologically-driven features of the modern cigarette fabrication process, and recent years have seen similar innovations that streamline the growing and marketing of the tobacco crop itself. The many examples would include the introduction of machines to assist with reaping, automated regulation of temperature and humidity in bulk curers, and new and efficient ways to transport tobacco bales at the auction floors. This dissertation will describe a possible application of a relatively new technology, namely *machine vision*, to a stage of the tobacco production process known as *grading*. The author (who is a non-smoker) hopes that this work will highlight a production process which is still in many ways a hostage to fashion and subjectivity, and yet which satisfies the demands of and offers employment to a very large number of people today.

1.2 Economic perspective

Total annual world tobacco production fluctuates considerably from year to year, depending upon such factors as the success of the global harvest, world stock levels from the previous season and where those stocks are held, and the anticipated price, especially in seasons following a poor crop in a major producer such as China. Table 1.1, which gives the total production figures over the last six years, reveals a variation between the extreme years of 2.029 million tonnes compared with a mean value of 6.739 million tonnes, even over this short period of less than a decade. This volatility in the total production figures is offset by the fact that world stocks tend to amount to just under a full year's production, although it must be noted that China holds a very large percentage of the stock. Table 1.1 also includes the estimated stock holdings as they have stood over the past few years, both including and excluding the stock held by China [102].

The cropped cultivar and the treatment by producers of tobacco also have great worldwide variation, with the result that total world production may be divided into several very distinct types, each of which is most suitable for the manufacture of a specific style of smoking material. In 1929, the Bureau of Agricultural Economics of the United States Department of Agriculture (USDA) identified six *classes* of tobacco, comprising a total of 26 *types* [30], each distinguished by the tobacco variety or cul-

tivar used, or by the process of curing the leaves, or by the eventual end use of the tobacco once sold (e.g. cigars, cigarettes, etc.). Many further types, grown in countries outside the United States, were also identified in this classification. Fortunately from the point of view of simplicity, the overwhelming majority of modern tobacco production belongs to one of seven of the types identified, which are listed in table 1.2, together with the approximate total of each type produced in 1997 [103].

Both the production and the consumption of tobacco and its products exhibit some marked regional trends. Whilst wordwide production grew at about 1.8% per annum (averaged over the period 1974-86), this was represented by a mean annual growth rate of 3.4% in developing countries and a mean annual decline, -0.9%, in the developed world. Countries like the United States, Canada, Argentina, Turkey and Japan, acting in response to the increased awareness of the dangers of smoking and in recognition of its costs to the public health sectors of their economies, have all cut back on production since 1975. Other countries, notably Brazil, Malawi, China and Indonesia, expanded production very significantly over the same period, with the latter three achieving mean annual growth rates of 6% to 8% for over a decade after 1975. Total world consumption in 1974-86 rose by an average of 2.4% per annum, but this was represented by a mean rise of 4.8% (or 1.9% per capita) in developing countries and a mean fall of -0.4% (or -1.5% per capita) in the developed world [28].

More recently, despite new recognition of the health threat posed by passive smoking and notwithstanding legislation in many developed countries that curbs advertising, discourages teenage smoking and reduces retail outlets, both worldwide production and consumption continue their upward trends, with leaf production expected to be

Total world tobacco production, 1993-1998			
Year	Total	Stock	Stock excl. China
	Mt	Mt (est.)	Mt (est.)
1993	7.917	7.7	5.3
1994	6.140	6.6	4.9
1995	5.888	5.7	4.3
1996	6.417	5.8	3.9
1997	6.961	6.1	4.1
1998	7.113	-	-

Table 1.1: **World tobacco production statistics in millions of tonnes of green leaf.**

Estimated breakdown of 1997 world production by the major tobacco types	
Type	Production (Mt)
Flue cured	4.500
Burley	0.940
Dark air or sun cured	0.930
Dark fire cured	0.048
Light air cured	0.097
Oriental	0.620

Table 1.2: **Millions of tonnes of green leaf of major types produced in 1997.**

growing at 1.9% (2.4% developing world, 0.8% developed) and tobacco consumption also to be growing at 1.9% (3% developing world, -0.1% developed) in the year 2000 [36]. Price increases through levied taxations, the impositions of higher penalties and payments on the players in the industry, the threatened listing of tobacco as a drug and, most recently, the negotiations for a massive (US$246 billion) reparatory legal settlement [92, 93] in the United States, are curbing (but not strongly reducing) the demand for tobacco products in countries such as the United States and United Kingdom. This is more than offset, however, by the opening up of new mass markets in formerly-communist Eastern Europe, developing Africa, Asia and South America.

While United States domestic cigarette consumption has fallen by 2% - 3% in most years since 1988 [104], the United States is still a massive (and growing) [101] importer of cured tobacco, which it processes and then re-exports for use in other countries. As an example, United States tobacco imports soared from 1995 to 1996 by over 60% [76], and United States cigarette sales to Eastern Europe increased 5-fold from 3.7 billion pieces to 19 billion pieces in the same short period of time [101].

In contrast, China, which is by far the world's largest tobacco producer, exports only a very small percentage of its crop, the bulk being accounted for in domestic usage. Table 1.3 lists the world's main tobacco producing nations and gives, for the five largest exporters, the dry weight of their exportation of the main type of cigarette-manufacturing tobacco, which is the flue-cured product [103, 101].

This dissertation will focus almost exclusively on the classification of flue-cured tobacco, because this is the dominant style grown in the world today, as Table 1.2 shows,

The main tobacco producing nations and flue-cured tobacco exporters		
Nation	1997 Total Production	1998 Flue-Cured Export
	1000 tonnes	1000 tonnes dry weight
China	2400	80
USA	433	108
Brazil	430	230
Zimbabwe	196	175
India	193	75
Argentina	75.5	
Canada	70	
Tanzania	35	
Other	657	

Table 1.3: **Estimated production and flue-cured export of the main producers**

and because of its central importance for the cigarette industry. It will also concentrate on the tobacco grading practices in Zimbabwe, which, as Table 1.3 reveals, is the world's fourth largest tobacco producing country, and the second largest flue-cured tobacco exporter.

The high cash-crop value per unit of land, the opportunity which it offers for substantial rural employment, and its foreign exchange earnings potential all make tobacco a favoured crop in a developing country whose climate will support its cultivation. The geography, soil conditions and normal seasonal rainfall in parts of the north and east of Zimbabwe are ideal for the farming of the Virginia tobacco crops which, when flue-cured, are in highest demand for cigarette manufacture. Moreover, Zimbabwean rural labour has hitherto been in good supply, with minimum agricultural wages set by the Government at significant yet not burdensome levels for the entrepreneur.

Thus, for many years Zimbabwe has accounted for about 1.8% of total world tobacco production, or about 4% of the total global flue-cured crop [81]. Local consumption is very small, and the great majority of the Zimbabwean flue-cured crop is exported. The tobacco industry in Zimbabwe is not only the nation's main employer, but also its largest earner of foreign exchange. In 1996, tobacco accounted for 30% of Zimbabwe's total exports of Z$24.2 billion [27], while in 1997 160 million kg of flue-cured and burley tobacco were exported, earning the country Z$6.6 billion [37].

Some 98% of production is exported, with the remaining 2% representing domestic consumption and wastage losses that are incurred in pre-export processing.

The concentration of the Zimbabwean economy on the production of high-quality export tobacco, and the consequent pressure on the Zimbabwean tobacco industry to meet the highest standards in every respect, have long been noted. Akehurst [3], for example, observed that

> Zimbabwean tobacco growing is unusual, as an agricultural industry, in that a large export trade has been built up on a relatively small home demand. In 1959 exports were nearly 80 per cent of production and the figure has since risen because production has increased considerably against relatively static local consumption. It is thus exceedingly vulnerable and must operate on the strictest criteria, in order consistently to produce, in every respect, the standards desirable by its markets

and also that

> The Zimbabwean industry has no equal in the world for the degree with which all sectors involved, farmers, merchants, manufacturers and the research organisations are effectively pulling together in the same direction, towards lower-cost profitable production of leaf styles covering the requirements of the world's main importing countries.

Zimbabwean flue-cured tobacco is carefully produced to the highest world standards and therefore provides an excellent example by which to assess industrial techniques such as the grading criteria that will be considered here. The findings of this dissertation, made on the basis of a study of typical Zimbabwean flue-cured tobacco, may confidently be extended to flue-cured tobacco as it is grown in all other parts of the world precisely because of Zimbabwean success in maintaining stringent international standards.

1.3 Problem statement and scope of study

Accurate grading of flue-cured tobacco is an essential prerequisite to selling it [80]. Grading is currently performed manually in a process which, with a few exceptions,

has altered very little over the past century, and which remains very labour intensive. Although flue-cured tobacco grades are defined qualitatively and at some length in the literature that is available to farmers [90, 54, 60, 94], quantitative criteria for the grading of tobacco are scant or non-existent [1, 11].

The purpose of this study is thus two-fold. Firstly, it is an investigation into the possibility of using machine vision techniques to grade, or assist with the grading of, flue-cured tobacco leaves. It offers the theoretical basis for the design of an automated or semi-automated software system for grading, which may be faster, more accurate and cheaper to operate than current manual methods. Secondly, in formulating and testing image processing algorithms for the automated grading of tobacco, this study devises quantitative criteria which achieve virtually the same classification success rates as human graders operating with their relatively ill-defined qualitative guidelines. This dissertation may enhance understanding of the grading process, and indeed improve its efficiency, by offering objective criteria for the grading of flue-cured tobacco leaves.

The study will be restricted to two of the major attributes by which leaves are graded — colour and plant position. Although for both of these (and particularly for plant position) a human grader will use some non-visual cues, such as the aroma or feel of the leaf, the assumption here is that these attributes may adequately be judged from only the visual information that is available in an image of the upper surface of the spread, flattened leaf. This dissertation aims to prove that, even with a much more limited set of inputs than is available to the human being (i.e. with visual inputs only), a machine vision system could be expected to achieve grading accuracies that are very close or even superior to those of typical human graders.

The practicalities of the hardware implementation of an automated leaf grading system will not be discussed, except insofar as they were issues in the practical component of this project (e.g. in the illumination of the leaves for imaging purposes). It is assumed that a fully implemented automatic grader could be faster than a human grader because of the mechanical properties of its moving parts and because of the processing speed of which its electronic hardware would be capable. This study will concentrate on the development and testing of image processing algorithms for leaf grading, but will not emphasise the coding or the optimisation for speed of these algorithms, except where this became an issue in the testing of the algorithms.

1.4 Literature review

Prior to 1926, published literature that describes the process of tobacco grading often serves today to emphasise how very subjective a process it was. An excellent treatise for its time, for example, is *A Textbook on Tobacco* [98], published in 1914, in which the author's description of the "sorting" of tobacco highlights this impression:

> The tobacco is usually divided, with infinite care and judgement, into the following kinds: Brown, dark gray, light gray, yellow, multicolored, coarse not speckled, slightly speckled, dark and brown slightly speckled, gray and light speckled all colors, little-broken dark and brown, little-broken gray and light, much-broken all colors, sweepings, and trash.

In January, 1926, pursuant to the Warehouse Act, the United States Department of Agriculture published [87] what the *Yearbook of Agriculture* of that year [88] described as "a complete and systematic classification of All American leaf tobacco". This was a huge step forward in tackling the bewildering taxonomy of tobacco types and sub-types already prevalent. It established the division of all tobaccos into *classes* (usually depending on their method of curing or their usage in cigars), each of which was subdivided into *types* (often by reference to the American state where it was grown). Within each type, a set of *grade groups* was then assigned, which depended on the leaf's position on the plant or on its general coloration, among other considerations. Although this classification scheme is clearly artificial, it has proved helpful in establishing a common reference for growers and buyers, and has been adopted in many countries worldwide. The USDA literature of the time identified classes and types by arabic numerals and the grades by the alphanumeric sequences (e.g. B2L, B4F, X1L etc.) that are still essentially with us today [90].

Since 1926, the literature relevant to tobacco grading, and particularly to the aims of this study, seems to fall into about eight categories: historical books, methodological publications, advisory magazine articles, statistical surveys, newspaper articles, technical papers, books devoted to tobacco or tobacco growing and, finally, textbooks on image processing and classification.

Historical books (e.g. [23, 42, 10, 47, 75, 73]) are valuable in setting the historical context of the need for tobacco grading and in highlighting the methods that were originally employed to sort tobacco by perceived type. It is clear, for instance, that

the buyers' tastes have always been of paramount importance to tobacco growers, and that the growers' preoccupation with packing their product into bales of uniform type has stemmed from the buyers' requirements in this regard [73].

Government, Growers' Association or Marketing Board publications [89, 21, 54], which cover the methodology of tobacco growing, curing, grading and sale, are extremely valuable in placing grading in its correct perspective within the flue-cured tobacco production process, and in defining exactly how grading should be carried out. Reading this literature often makes it clear that flue-cured Virginia tobacco is now handled and graded almost identically worldwide [60, 94], which gives confidence that an automated grading system with objective grading criteria would have wide applicability. Of particular value to understanding the grading of tobacco in Zimbabwe has been the *One Industry Approach to Grading and Presentation of Flue-Cured Tobacco* [80], which describes the modern classification scheme as applied to tobacco at auction. There are several references [21, 54] which cover the practicalities of grading leaf tobacco in minute detail, including giving recommendations for grading shed layout, leaf steam conditioning, lighting, humidity, the keeping of records, labour requirements, ergonomics, inspection and quality control, and rates of pay. The practical considerations arising from these references have an important impact on this study, especially as they relate to matters such as the illumination of the grading sheds or the volume of tobacco that a human grader can classify in a day.

The 1960s appear to have been a heyday for national tobacco growers' journals, and regular publications that have provided many articles of relevance to tobacco grading include the *Rhodesian Tobacco Journal*, the *Australian Tobacco Journal*, the *Canadian Tobacco Grower*, *The Bright Leaf* (an Ontario tobacco growers' journal) and *Farming in South Africa*. At a time when individual growers were bringing very innovative methods to the industry, these journals served to inform farmers of the latest practices. By reading back-issues of such publications, one can trace some important developments in grading techniques. In Canada, for instance, where tobacco was traditionally grown and graded by isolated farming families, grading was greatly assisted by installing motorised conveyor belts [17, 65, 18]. The introduction of artificial fluorescent lights extended the hours of day during which workers could grade tobacco without adversely affecting their grading accuracy [64, 71, 63]. Much more attention was paid to technical issues such as the layout of the grading shed [63], the quality of the graders' colour vision [67] and the work efficiency and remuneration scales

for grading [66, 69, 70]. Plentiful advice was also given to farmers as to simple systems for grading leaf into a relatively small number of categories prior to sale (e.g. [68, 19, 48]). From the point of view of this study, these developments are of interest because they represent injections of technological or economic innovation into a very labour-intensive activity, and, as such, they are forerunners of the kind of innovation which this study hopes to bring, four decades later, to the same process of leaf grading.

Statistical surveys and newspaper articles are of particular value in gleaning data that justifies the economic potential of bringing machine vision to an automated grading system. Certain issues of *Tobacco Trends* [101, 102, 103, 104], published quarterly by the Market Information Department of the Zimbabwe Tobacco Association, are very useful in explaining the current economic context of flue-cured tobacco production and marketing both internationally and in Zimbabwe itself. The more detailed *Flue Cured Crop Annual Statistical Reports* [81, 83], prepared by the Technical Division of the Tobacco Industry and Marketing Board (TIMB) in Zimbabwe, provide insight into the annual earnings from tobacco in Zimbabwe and, most importantly for this study, have useful breakdowns of each annual crop into the top 100 grades sold, listed by percentage of the crop and price at auction. The breakdowns help one to assess the value of accurate grading and, indeed, to see certain cases in which there is no significant difference between the prices of two different grades, even though the leaves of the two grades differ markedly in appearance. The TIMB's weekly *Flue-Cured Tobacco Market Report* (e.g. [79]) and their *Statistical Summary of Flue-Cured Virginia Auction Sales from 1936 to Date* [78] are also both fertile sources of useful statistics.

There is so far very little published work that deals formally with the possible applications of machine vision to the grading of leaf tobacco. Campbell [11] points out that both farmers and buyers need to be aware of changing fashions in the consumer market, and that cured tobacco will continue to be purchased with regard to certain subjective selection ideals, even though acceptable objective measures of quality (such as chemical content) must increasingly be brought to bear in the trade. He makes the important points that tobacco, in contrast to many other agricultural products, "has to meet virtually no ... objective specifications prior to purchase" and that "even colour has never been defined by the USDA". This study has the opportunity to rectify this, especially in terms of colour but also with reference to the measurable features of leaf quality and plant position.

Abdallah [1], writing 30 years ago, commented that " it is still rather difficult to establish a satisfactory means of accurately measuring elements of quality of leaf tobacco, i.e. body, texture, thickness, color ... " and that "[t]he need for objective methods of defining and measuring quality is the most important area of research in tobacco today". Whilst there has been much progress since these comments were made in measuring the chemical content of leaves [16], it is the view of the present author that a satisfactory means for objectively defining and rapidly assessing leaf colour and shape is only now becoming available through digital image processing techniques.

Some work in this direction has already been undertaken. For example, Tucker and Chakrabarty [85] have recently produced software that uses image segmentation and classification techniques to identify blight and rust lesions on oats and sunflower leaves for the purpose of rapid disease assessment in the field. There has apparently been no such initiative to date with tobacco leaves, but it would seem an attractive area of study as it offers the possibility of rapid disease detection and possibly even of automated diagnosis. For tobacco, there has been some research into the estimation of leaf area in Japanese burley [51] and also into a Cuban dark tobacco variety called *corojo*[6]. Both of these projects conclude that leaf area can be estimated with only small errors, in the case of the cultivar they were using, from simply-extracted features such as leaf length and width. Such feature extraction is what this study seeks to automate by applying machine vision techniques to the classification of leaf shape, in order to estimate the plant position from which the leaf was reaped.

Research into the quantitative assessment of tobacco colour has concentrated on the use of direct visual comparisons of the tobacco leaf with a physical colour scale or chart. Shiga *et al* [61] developed a colour plate which they used to assess both burley and a Japanese cultivar called *shiroensu*. Using their plate they found that they were able, under a wide range of lighting conditions, to assess green leaf colour very precisely, and from it to identify chlorophyll content with high confidence. In a similar experiment, Akimoto *et al* [4] found that the use of a colour scale enabled them to determine the readiness for reaping of the leaves of a particular tobacco cultivar and that, by bringing reaping forward by as much as a week, the use of the chart could lead to an improvement in reaped leaf quality. Of particular relevance to the current project is a paper by Carotenuto [12] who used visual comparison with Munsell colour charts [50] for the colour assessment of Italian flue-cured Virginia tobacco, and who showed that this technique was not only valuable for the assessment of the market grade of

the tobacco but could also be used to derive quantitative estimates for leaf reducing sugar, total alkaloid and total N content. These papers all seem to point to the hitherto unexplored area of automated leaf colour and shape determination, because they each substantiate the fact that quantitative assessment of leaf colour and size is possible. The application of computer-based digital image processing techniques to this assessment is just the next logical step in making the assessment more quantitative, more objective, more consistent over time, faster and less labour-intensive.

1.5 Summary of procedure and dissertation preview

This dissertation aims to produce digital image processing algorithms which are effective in the grading of flue-cured tobacco. This means that the algorithms must grade flue-cured leaf into widely-used standard categories [90, 80] with a misclassification rate that is comparable with (hopefully equal to or less than) the misclassification rate of an average human grader (which is about 10%). The algorithm will operate on single images of the tobacco leaves which are to be graded, and human graders will have graded the same leaves in a process which must be "blind" as far as the algorithm is concerned. In one conceivable method of comparison of grading performance, the human graders may have access only to the leaf images, exactly as the computer algorithm does. Whilst this may be thought of as a *fair* comparison of grading accuracy, this dissertation aims to produce an *effective* machine vision based grading algorithm, and so the human graders used in this project were experts who had had full access to every imaged leaf (touch, sight and aroma) while grading it.

The method of procedure in this project is summarised in Figure 1.1. Given the statement of the problem and definition of the aim of the research, as stated in section 1.3, the next step was to acquire some suitable tobacco leaves. The selected leaves were all taken from bales which had been successfully sold at auction on the Tobacco Sales Floors in Harare, Zimbabwe. Grading of the whole bales had therefore been done by official graders at the auction floors, and had been accepted as correct by both the seller and the purchaser. Nevertheless, arrangements were made for each leaf to be re-examined individually by an expert grader, who assigned it a grade which was recorded for the purposes of this project. Each leaf was then prepared and photographed under suitable and stringently consistent conditions of lighting and scale. Later, each photographic image was digitised for use in the computer classifier algo-

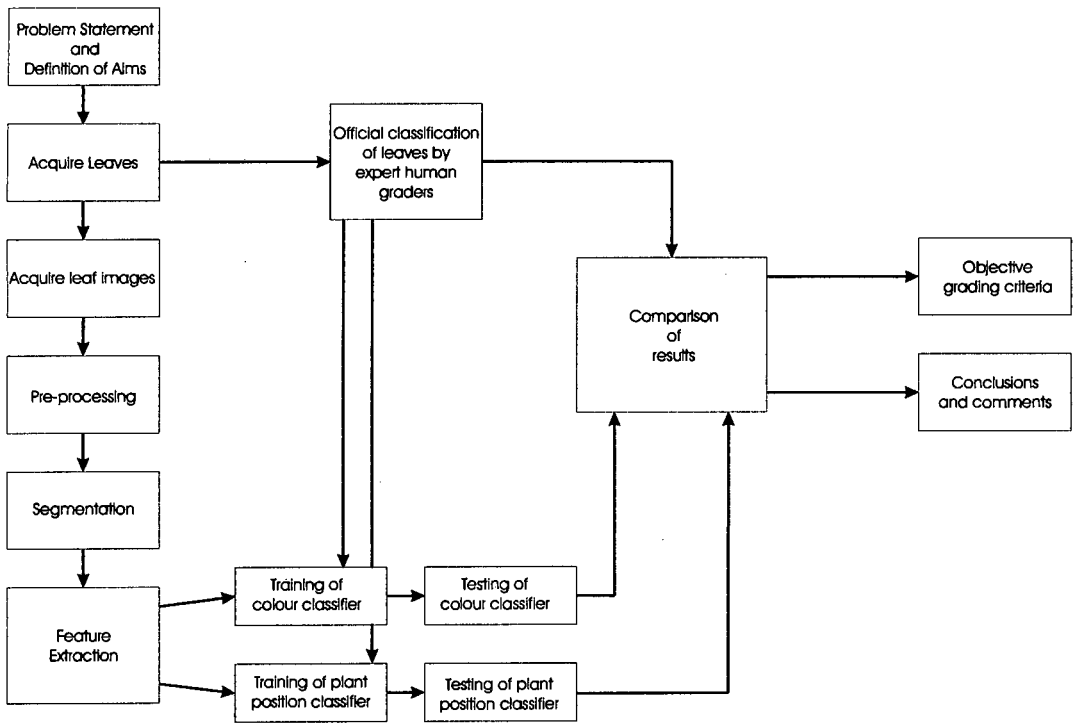


Figure 1.1: **Summary method of procedure for this project**

rithms. To ensure the highest standards of consistency, images were pre-processed prior to the isolation (segmentation) of the main objects in each image. Once each object had been unambiguously recognised, features could be extracted from it. Such features may be as simple to extract as, say, the length or pixel area of the segmented object, or they may be the results of quite extensive mathematical derivation. Whichever the case, the process of choosing appropriate features for a given classification purpose is notoriously heuristic and often *ad hoc* [32], and this dissertation will cover in detail the reasons behind the feature choices that were made in this project.

Two classifier algorithms were developed, one for classifying tobacco leaves by colour, and one which deduced the plant position at which the leaf had been reaped from the stalk of the tobacco plant. Each classifier was trained by working with a subset of the available leaf images, and was then tested on another subset of leaf images which was quite distinct from the training set. From this, each classifier earned a "score" of the percentage of correctly-classified images, which was derived by comparing the classifier results with the grades given to each leaf by the human expert grader. Arising from these results, it has been possible to give some objective criteria for the grading of tobacco leaves by colour and plant position, and to draw certain conclusions

regarding the feasibility of the machine vision grading of flue-cured tobacco leaves.

The first (present) chapter of this dissertation provides an introductory statement of the problem of grading tobacco, and sets this problem in its historical, economic and intellectual contexts.

Chapter 2 deals in detail with the way that tobacco is graded, giving a full explanation of the grades in use in Zimbabwe and around the world and discussing the possible advantages of automating the process. Innovations in grading over the last 40 years are briefly mentioned, and the proposed introduction of machine vision is viewed in the context of the on-going application of technology to the grading process.

Chapter 3 is a review of the concepts in digital image processing that were used in the development of machine vision algorithms for this project. The computer-based characterisations of colour and shape, which are central to this study, are explained in some detail, and several methods and transforms that were considered or accepted for use in the grading algorithms are introduced.

Chapter 4 explains how photographic images of actual tobacco leaves were acquired, and describes the digitisation and pre-processing of the images for use in the classifier algorithms.

Chapter 5 presents the theory of classification on the basis of features extracted from digital images. The chapter goes on to describe how the choice of features selected for each leaf may be improved by means of discriminant analysis.

The classification of tobacco leaves by colour is then presented in Chapter 6, which discusses the extraction of suitable features for this purpose, the results achieved, error rates and the setting of objective criteria for colour classification.

The classification of tobacco leaves by plant position is handled in a similar way in Chapter 7, where a different set of features is selected for the classifier. Results are again presented and discussed, and the quantitative variation of leaf shape with plant position in typical plants is described.

Finally, in chapter 8, a summary is given of what is felt to have been achieved by the project, and the main results are listed along with suggestions for further investigations in this area.

Chapter 2

The Grading of Flue-Cured Leaf Tobacco

2.1 Plant position and colour

So-called "bright" Virginia tobacco grows as a large, somewhat leathery plant (see figure 2.1). At maturity, it may attain a height of about 90-150cm (after topping), and its approximately 20 leaves change colour from a dark green to a much lighter,

Figure 2.1: **Tobacco plants growing in the field**

brilliant green, sometimes with yellowish overtones. More leaves would grow, and the plant would grow higher and indeed produce flowers, but removal in the land of the main budding shoots (known as *topping*) arrests the progress of the plant's upward growth so that the main growing effort is redirected to the leaves [3]. Leaf shape and size both vary as functions of position on the plant's main stalk, and many leaves are enormous, with lengths in excess of 70cm being not uncommon. The leaves closest to the ground ripen first and must be reaped while the upper leaves are still growing (see figure 2.2).

Figure 2.2: **Partially reaped tobacco**

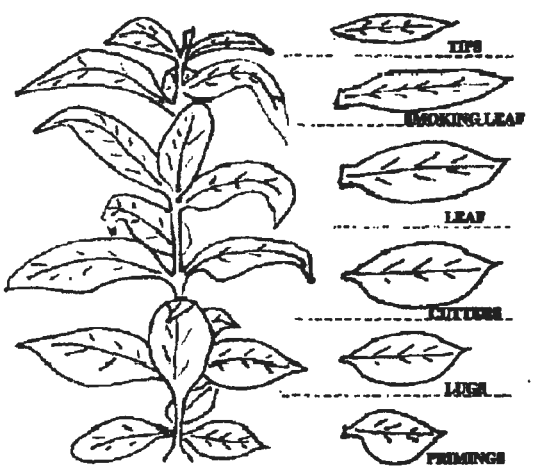


Figure 2.3: **Virginia tobacco plant positions**

There are other significant differences between the lower and upper leaves of a Virginia tobacco plant. Figure 2.3 shows the relative shapes of leaves and gives the names of the six *plant positions* (or *groups*) that are sometimes referred to in this type of tobacco. Primings and lugs, closest to the ground, tend to be of thin lamina and quite narrow midrib, whereas the upper leaves and tips are more thick textured, with prominent midrib and veins. A very significant chemical difference is that the upper leaves are much richer in nicotine that those in lower plant positions. There are also differences in the oiliness and aroma of the leaves in the different plant positions (and consequently in their

flavour when smoked). These considerations make it essential for cigarette manufacturers to purchase tobacco in bales which contain leaves of one plant position or group only. At a later stage, the manufacturer can then blend leaves from different parts of the plant to achieve a mixture that gives a particular smoking flavour, nicotine content, consistency, ash characteristic, etc. It should be noted that pictures like figure 2.3 and detailed qualitative descriptions of leaf variation with plant position are a typical part of what is currently used to train a person to grade tobacco, and that with experience a human grader can become quite proficient at recognising the plant position of a cured leaf, having seen, felt and smelled it.

(a) Before (b) After

Figure 2.4: **Tobacco coloration before and after curing**

Once reaped, tobacco leaves are taken immediately (and with much care to avoid breaking or bruising them) to large barns where they are then hung up and *cured*. Curing is a treatment of the leaves in bulk which alters and then stabilises their biochemical properties, so allowing them to be stored for many months thereafter and resulting in a product which is acceptable to smoke. There are several methods of curing tobacco including hanging it in the open air, laying it under the sun, and treating it with smoke; and each of these is traditionally used with certain types of tobacco in various

parts of the world. Virginia "bright" tobacco is most usually *flue-cured*, which means that great numbers of leaves are hung together in high barns through which heated air is pumped. The air rises from the bottom of the barns and eventually escapes through vents at the top. Barn conditions (temperature and humidity) are carefully controlled over a period of several days, at the end of which the leaf will have fermented and so attained one of a range of brownish colours, a recognisable tobacco aroma, and a firm, somewhat pliable, texture. Curing tunnels and bulk curers are increasingly taking the place of traditional barns, but operate on exactly the same principle of passing heated air over the drying leaves to accelerate and control fermentation. Figure 2.4 shows a barn of tobacco both before and after curing, illustrating the colour change that takes place in the leaves.

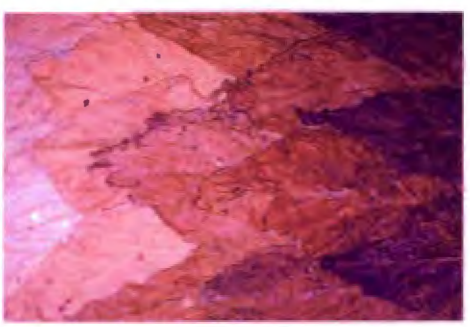


Figure 2.5: **Tobacco colour variation**

The colour of a cured leaf depends on the type and quantity of certain fermentation products within it, and this in turn may relate to the type of tobacco plant from which the leaf came, the way that the plant was treated when in the land, the climatic conditions that prevailed when it was growing, or the plant position from which the leaf was reaped. Further discussion of how leaf colour is achieved is beyond the scope of this dissertation, but the crucial outcome is that the cured leaf colour of tobacco of identical type may vary dramatically from farm to farm or from reaping to reaping on a particular farm, and is indicative of important differences in the smoking characteristics of the leaf. Figure 2.5 shows a collection of several leaves arranged to illustrate this colour variation.

Flue-cured Virginia tobacco leaves are classified into one of five colours named pale lemon, lemon, orange, light mahogany and dark mahogany. Figure 2.6 shows five cured leaves, each of which is typical of its colour class. Again, it may be noted that human graders will usually learn to recognise tobacco colours through their long experience of handling leaves and with reference to written descriptions of the colours. They do not presently have access to quantitative descriptions of the leaf colours, nor even to pictures like figure 2.6. Leaf colour is somewhat easier to judge than plant position but, even so, a human grader may experience difficulty in a marginal case. As with plant position, the importance of correctly sorting tobacco leaves by

Figure 2.6: **Pale lemon, lemon, orange, light mahogany and dark mahogany**

their colour is that different colours of leaves have different concentrations of the compounds which affect the smoking experience. In particular, the darker-coloured leaves are richer in nicotine, and a manufacturer will wish to blend leaves with this sort of consideration in mind.

2.2 Quality

Apart from plant position and colour, leaves are also graded according to their *quality*. It is possible for a ripe leaf to endure the growing, reaping, transportation and handling processes prior to curing and then to be cured entirely successfully so that it emerges from the barn as an essentially unblemished, highly desirable cured leaf. In this case its quality is describes as "choice" or "very good", and it will be given a quality grade of 1. It is more common, however, for some degree of damage to occur to a leaf during its history. Attacks by small animals or insects or heavy rain or hail may damage the lamina of the leaf while it is still on the plant. Diseases, some quite harmless to the leaf and others very malevolent, may likewise leave their mark. Poor weather conditions (such as drought or flood during growth) may also be detectable, as may errors in the management of the plant (such as deficiencies or excesses of nutrients,

herbicides or pesticides). Careless reaping or transportation to the barns of vulnerable ripe leaves can lead to bruised or broken areas which will become dark and shattery as a result of the curing process. Furthermore, if the temperature and humidity of the barn during curing are not meticulously controlled, leaves may manifest scorching or watermarking, or they may retain some of their original green coloration or, worst of all, the whole barn full of leaves may begin to mould or rot. Very ripe leaf is most desirable from the point of view of buyers, but is most at risk to many of these types of damage. On the other hand, leaf which is reaped when immature will be close-grained and "slick", and this will be a point against it at auction. It is small wonder, then, that the curing period is a very anxious time for the farmer!

(a) Quality 1 (b) Quality 3 (c) Quality 5

Figure 2.7: **Cured tobacco leaves of varying *quality***

Figure 2.7 shows a leaf of quality 1, together with two leaves *of the same plant position and colour* that have suffered some forms of damage, and have been graded as quality 3 and quality 5. What is immediately apparent is that the degradation in the lower quality leaves has left dark blemishes on them that give the entire leaf a darker appearance. The leaves in figure 2.7 are all lemon in colour, and indeed they all have areas of undamaged lamina which manifest the true lemon colour, but the lower qual-

ity leaves, and especially the leaf of quality 5, may appear to an inexperienced eye to belong to the orange or even to the light mahogany category, simply because of the density of blemishes on their surfaces. Human graders learn to judge quality and colour separately, but the impact of the quality upon the apparent colour of a cured leaf is an issue to which this dissertation must return in considering the grading of leaf colour by machine vision.

2.3 Official grading schemes

In 1926, the United States Department of Agriculture (USDA) specified in great detail [90] how Virginia flue-cured tobacco should be graded. Their rules for grading define a *grade* simply as " a subdivision of a type according to its group, quality and color", where the *group* of the tobacco means either its plant position or else the consistency of the tobacco body if it is in the form of strips or scrap. A code letter is assigned to each group, a number to each quality, and another letter to each possible colour, as shown in table 2.1.

Table of standard grade factors defined in the USDA's *Official Standard Grades for Flue-Cured Tobacco* [90]

GROUP		QUALITY		COLOUR	
Name	Code	Name	Code	Name	Code
Primings	P	Choice	1	Whitish Lemon	LL
Lugs	X	Fine	2	Lemon	L
Cutters	C	Good	3	Orange	F
Leaf	B	Fair	4	Red	R
Smoking Leaf	H	Low	5	Greenish	V
		Poor	6	Variegated	K

Table 2.1: **Grade identifier symbols used in the United States**

Thus, an orange leaf of good quality reaped from the "leaf" position on the stalk of the plant will be assigned the *grade mark B3F*. Up to two further letters may be added to this 3-symbol grade mark if there is a need to indicate any extra feature of the tobacco such as darkness (*D*), green (*G*), oxidation (*O*), mixed grading (*M*), etc. Each possible grade combination is carefully described in the official publication [90] but, apart from

measures of length, percentage uniformity, percentage injury tolerance and percentage wastage, all of the descriptors are qualitative in nature. It is evident that many years of experience would be required to become an expert in grading using this system, and that grading speed combined with accuracy would, even then, be difficult to achieve. The total number of possible flue-cured Virginia grades in the American system is 153. "Grading" itself is rather loosely defined by the USDA [86] as "classification of products by standards of uniformity, size, freedom from blemish or disease, fineness, quality, etc.", but there are 29 rules and 85 subsidiary definitions given in the standards document which between them constitute a description of how to do it. An impression of the subjective and relative nature of the descriptions given for each grade may be gained from a quoted example [90]:

> C4KR Fair Quality Variegated Red or Scorched Cutters
>
> Ripe, open leaf structure, medium body, lean in oil, moderate color intensity, normal width, 16 inches or over in length. Uniformity, 70 percent; injury tolerance, 20 percent, of which not over 5 percent may be waste.

Several writers have attempted formal definitions of grading. Campbell [11], for instance, speaks of "sorting and preparing flue- or air-cured leaf for sale according to plant position, colour, maturity, and degree of blemish or damage to produce uniform lots", whilst Jahn [39] gives "GRADING: the selection and grouping of leaf according to its quality, color, thickness, elasticity, injury, etc." Akehurst [3], working more from the perpective of end *usability*, points out that graded lots ought "to be uniform within themselves, but distinct for some ulterior purpose or specification, such as smoking or manufacturing qualities". Another definition, given by the Tobacco Industry Marketing Board in Zimbabwe [80], focuses on the immediate need of farmers to sell their crop: "the aim of grading is to present tobacco in a manner that enhances both its use and value to the buyers by sorting out leaves of similar characteristics into uniform lots for sale". A definition which encompasses all of these views quite nicely, and which stresses plant position, appears in Voges' *Tobacco Encyclopedia* [95]:

> The sorting of cured leaves into lots which are, for their manufacturing purpose, homogeneous according to plant position styles and external appreciation. Factors to be considered include overall colour, blemish, damage, texture, leaf length and ripeness. These will be applied to varying

degrees, depending on the tobacco type and market requirements, to each plant position category.

Armed with plentiful advice about *how* to grade, tobacco growers worldwide have quite understandably been happy to live without an agreed formal definition of grading, and have converged on grading conventions that are all very close to the USDA's official system. Thus manuals on grading from as far afield as Ontario, Victoria Province in Australia [94] and Thailand [60] all talk in terms of plant position, quality and colour as the main determinants of grade, with extra features such as leaf body, maturity, texture, solidity, length, oiliness, elasticity, finish, width, grain and strength also being taken into consideration.

In the case of Zimbabwean tobacco grading, Campbell [11] takes the view that the USDA grading system (which was developed at a time when cigarettes were still a minority destination for graded tobacco) has lost some of its applicability because modern tobacco products favour slightly different leaf characteristics from those in vogue 70 years ago. Also, unlike in the United States, Zimbabwean grades are not linked by statute to the prices paid, and so the buyers can communicate their needs directly to the growers through the amount that they bid for a particular grade. In several ways, therefore, the Zimbabwe tobacco market is both more grade-conscious and more up-to-date than the other big markets worldwide, and this further motivates and justifies the use of Zimbabwean tobacco in this project. Official classification and grade markings are no less minutely defined in Zimbabwe than elsewhere, and table 2.2 shows the standard symbols that are used in Zimbabwe, and which will be the system referred to constantly throughout the rest of this dissertation.

In addition to the standard symbols shown in figure 2.2, Zimbabwean grading may employ one of three *styles factors*, denoting extra ripeness (*F*), immaturity (*K*) or a blemish called *guineafowl spot* (*Y*), and one of five *extra factors* indicating distinctive but benevolent spot (*A*), harsh nature (*D*), barn scorching (*Q*), temporary greenishness (*V*) or a set green colour (*G*). When used in combination with the 3-symbol standard grade mark, the style and extra factors may give a final grade that requires either four or five symbols for a full description of the tobacco [80, 82].

When one considers that *N. tabacum* is only one of two *species* of smokable tobacco, that flue-cured tobacco is only one of seven *classes* of this species defined by the USDA, that Zimbabwean Virginia tobacco is only one of several dozen identified *types*

Table of grades and grade symbols used in the classification of Zimbabwean flue-cured tobacco [77]					
GROUP		QUALITY		COLOUR	
Name	Code	Name	Code	Name	Code
Primings	P	Very Good	1	Pale Lemon	E
Lugs	X	Good	2	Lemon	L
Cutters	C	Fair	3	Orange	O
Leaf	L	Low	4	Light Mahogany	R
Smoking Leaf	H	Poor	5	Dark Mahogany	S
Tips	T	Poorest	NG		

Table 2.2: **Grade identifier symbols used in Zimbabwe**

of this class, and that the number of possible *grades* of this type implied by the 5-symbol full grading appears to be

$$6 \text{ groups} \times 5 \text{ qualities} \times 5 \text{ colours} \times 3 \text{ style factors} \times 5 \text{ extra factors} = 2250,$$

one is tempted to the view that that this is minute classification taken to absurd extremes! The bewildering proliferation of grades of US tobacco was noted by Tilley [75] 50 years ago, and the period since then has certainly seen no reduction or rationalisation of the official grading scheme. Indeed, in 1981 Akehurst [3] made the telling comment that "the reality in [the] commercial characteristics made by relatively small differences in leaf appearance is questionable".

There are, however, certain grade mark combinations that can never be seen in real tobacco. Lower leaves never achieve dark coloration when cured, so, for instance, mahogany primings do not occur. Furthermore, there are style factors which are never found in combination. Table 2.3, following the Zimbabwean Tobacco Industry Marketing Board's grading guidelines [77], shows the allowed grade mark combinations for Zimbabwean tobacco, including those for strip tobacco (A) and scrap (B), just for completeness.

There are four further restrictions in Zimbabwean grading, as follows:

1. F and G cannot appear in combination;
2. Y can only combine with D;

Group	Symbol	Qualities	Colours	Style Factors	Extra Factors
Primings	P	1,2,3,4,5	E,L,O	F,K,Y	A,D,Q,V,G
Lugs	X	1,2,3,4,5	E,L,O	F,K,Y	A,D,Q,V,G
Cutters	C	1,2,3,4,5	E,L,O	F	A,V
Leaf	L	1,2,3,4,5	E,L,O,R,S	F,K	A,D,Q,V,G
Smoking Leaf	H	1,2,3,4,5	L,O,R	-	-
Tips	T	1,2,3	L,O,R,S	F,K	A,D,Q,V,G
Strip	A	1,2,3	E,L,O,R	F,K,Y	A,D,Q,V,G
Scrap	B	1,2,3	-	-	-

Table 2.3: **Permitted Zimbabwean tobacco grade mark combinations**

3. In cutters (C), A and V cannot combine with F, so all cutter grades have three or four symbols, never five;

4. Choice tobacco (quality 1) may not combine with:
   (a) Y,
   (b) D, Q or G,
   (c) the combination KV.

Using all of this information, table 2.4 presents a calculation of the number of possible Zimbabwean grades, which is found to be 990 leaf grades plus 155 for strip and scrap, totalling 1145. This dissertation will not be considering the grading of tobacco by style factor or extra factor, but instead will be of relevance to the 3-symbol *standard grade mark* for leaf tobacco, for which there are 97 possibilities. Even restricted, as it is, to a study of the machine vision grading of leaves by group (plant position)

Permitted and forbidden grade factor combinations in Zimbabwean tobacco grading																								
GROUP→	PRIMINGS			LUGS			CUTTERS			LEAF			S. LEAF			TIPS			STRIP			SCRAP		
No of symbols→	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5
Possible	15	120	225	15	120	225	15	45	30	25	175	250	15	0	0	12	84	120	12	96	180	3	0	0
LESS by 1.			15			15						25						12			12			
LESS by 2.			60			60															48			
LESS by 3.									30															
LESS by 4(a)		3	15		3	15														4	20			
LESS by 4(b)		9	27		9	27					15	30					36	72		12	36			
LESS by 4(c)			3			3						5						12			4			
Total	15	108	105	15	108	105	15	45	0	25	160	190	15	0	0	12	48	24	12	80	60	3	0	0
Group Total	228			228			60			375			15			84			152			3		
GRAND TOTAL	990 leaf grades + 155 strip or scrap = **1145 possible grades.**																							

Table 2.4: **Number of permitted Zimbabwean tobacco grade mark combinations**

and colour (i.e. not taking quality into account), this project seeks to classify the leaves into one of 21 permissible categories. This is about the maximum number of categories into which human graders working on a particular farm would "rough grade" the crop prior to sending it to auction [70, 48, 55], and in practice the number of farm grading categories during a particular reaping would be more like 10-12 [66]. Hence, machine vision classification based only on plant position and colour may achieve a grading resolution which is as fine as that attempted at present by farm graders. Furthermore, farm methods of rough grading tend to grade tobacco leaves by reference to local, idiosyncratic categories that can be remembered by the (largely uneducated) manual labourers who do the grading work. Machine vision methods which would work by reference to the official grading standards may therefore have the potential to grade more accurately in the eyes of buyers, the immense importance of which to the financial success of farmers has often been stressed (e.g. [31, 84]).

2.4 The case for automation

Figure 2.8: **A farm grading shed**

Every tobacco leaf is graded twice before it is sold - once by the grower and his labourers prior to baling his crop (see figure 2.8), and once at the auction floor to establish its official grade mark for the information of potential buyers. The second grading may result in any of the 1145 grade marks outlined in table 2.4, and takes an expert grader about 30 seconds to assess by quickly sampling a few leaves from a bale

of tobacco. Farm grading, on the other hand, is a massive, labour-intensive process that may take several dozen workers on a typical farm, grading at the standard rate of $8\frac{1}{2}$ hours a day [54], six or seven months to complete each year. The work involves opening up each leaf with a quick unrolling action of the thumbs, inspecting its upper surface very briefly whilst gauging its feel, and then adding it to one of about a dozen heaps, as shown in figure 2.9.

In an $8\frac{1}{2}$ hour day, a grader is expected to grade about 10 000 leaves in this way, a statistic which highlights the tedious and repetitive nature of the task. It has been estimated [74] that in Zimbabwe alone some 20 *billion* leaves are individually hand graded each year. Grading sheds are hot, with inside temperatures frequently in excess of 40°C, and are kept very humid by pumping in steam in order to keep leaves in their optimal pliable condition. The current (September 1999) minimum wage for a grading shed labourer (not grading tobacco) is Z$1000.43, based upon a working month of 22 days. A grader of tobacco earns a minimum of Z$1038.55, and a foreman or grading supervisor would receive Z$1205.28 at the very least. These are gazetted minima stipulated by the government, and some workers (but by no means all) may receive a little more. These wage levels are seen in their harshest perspective when viewed against the current exchange rates of Z$6.25 = 1 South African Rand and Z$37.95 = 1 United States Dollar with the backdrop of an estimated annual inflation rate of 65%. The legislation specifies a maximum working week of 54 hours, but in practice most labourers will work for 45-48 hours in a typical week. It has been estimated that 258 000 people (140 000 permanent and 118 000 seasonal workers) were employed in the Zimbabwean tobacco industry in 1998, and of these about 35 000 were employed as graders from time to time. It is a sad fact that many of these people live very close to the line of absolute poverty.

Figure 2.9: **Grading tobacco by hand**

In this respect, automation is not yet a practical financial solution for an emerging economy in which labour is still cheap and in which many workers rely on their grading job for income. This is especially true in Zimbabwe, where 70% of the labour force work in agriculture and where the unemployment rate was estimated (in 1994) as being at least 45% [100]. Nevertheless, the automation of tobacco grading should be an attractive option in more developed countries such as, for instance, Canada, where grading has traditionally been undertaken by isolated farming families who have often implemented innovations (such as the use of conveyors) to assist in this uncomfortable and monotonous task [17, 65, 18].

Figure 2.10: **Grading: a time-consuming matter**

Wherever farm grading is carried out, considerable thought must be applied to making it as efficient as possible, not only because it is a labour-intensive bottleneck in the tobacco production process, but also because even a slight failure to grade to the exacting standards of the tobacco buyers may lead to a farmer's crop being turned down at auction or else bought for a lower price than it would deserve if correctly sorted. In the 1958-9 season, a team of industrial consultants assisted the local Tobacco Association in preparing a *Work Study Manual* [54] which dealt, among many other things, with the ergonomics and the labour deployment issues involved in running a successful grading shed. Minute attention was given to every person's movements and rate

of work, and particular focus was brought to bear on the productivity of individual workers, their suggest rates of pay and (in that more illiberal age) even to financial penalties to be applied to any worker who was found to have fallen short of quota or to have made grading errors! Several articles were published which reported that the new management methods had cut grading costs, increased grading output and simplified the grading process [66, 69, 68], but grading remains even today a very time-consuming matter which might be achieved far more quickly, cheaply and accurately with the introduction of some degree of automation.

Chapter 3

Some Image Processing Concepts

3.1 Digital images

A digital image is a quantised two-dimensional array which represents the spatial distribution of light intensity, and also possibly of colour, in some actual object or scene. Numerous methods exist for acquiring digital images, but there are broadly two types of approach: either the image is captured directly from life using a digital imaging system such as a CCD-embedded camera, or else a pre-existing conventional

(a) 400×980 (b) 50×123 (c) 7×17

Figure 3.1: **The effect of varying pixel resolution in an image**

image (a photograph) may be scanned and *digitised*. In either case, the acquired digital image is quantised in two senses.

Firstly, the image will consist of an array or *raster* of spatially distinct small elements of equal size and shape, known as picture elements or *pixels*. The detail visible in the digital image will be limited by the size of the pixels in relation to the overall image dimensions, and this relationship may be used to express the *resolution* of the image. Figure 3.1 shows three digital images of the same tobacco leaf at different resolutions to illustrate this concept.

(a) 256 levels (b) 4 levels (c) 2 levels

Figure 3.2: **The effect of varying the number of grayscale levels in an image**

Secondly, each pixel will be associated with one of a fixed number (usually 256) of distinct light intensity levels. The impression that this gives, if the intensity values in the digital image are printed or otherwise displayed, is of a black-and-white photograph or monochrome television screen image; and if the pixels are sufficiently small and the number of intensity levels sufficiently large, then such a *grayscale* image may appear to the human eye as a good representation of the scene in reality. Figure 3.2 shows three grayscale images of a tobacco leaf each with a different number of intensity levels: one may note the presence of false contours in the image with only four grayscale levels. The impression of realism may be further greatly improved if each pixel is associated with a triplet of numbers in the range 0-255, each representing primary colour content, as described in the next section.

Although neither the spatial resolution nor the grayscale quantisation of today's digital images approaches those of which either conventional photography or human visual perception is capable, digital imaging nevertheless has several important virtues. By allowing a scene to be captured as an image which may be represented as a finite series of numbers (three numbers per pixel for a colour image), digital images are easily stored in modern computers, and may rapidly be retrieved from such storage or transmitted virtually anywhere worldwide through readily-available technology. Furthermore, systematic operations may be carried out on the array of stored numbers that represent an image, and this is known as *image processing*. As computer processors of ever-increasing speed continue to be produced, and as the times taken for storage, retrieval and mathematical operations continue to fall, image processing becomes ever-easier to implement as part of a decision-making sequence of programming steps (an *algorithm*) such as one might find within an automated process, for example. The speed and power of many image processing techniques, either in enhancing visual detail or in extracting detail automatically from images, are the principle motivations for integrating digital image processing algorithms into industrial systems. There is still a huge potential for research and development of industrial applications in what may be termed *machine vision*, and the world's tobacco industry is certainly no exception in this regard.

3.2 The specification of colour

A machine-vision system for the colour grading of tobacco must evidently reproduce the action of the human eye, at least to the extent of an ability to distinguish as well as a human does between the leaves of the five colour grades. Although there is no actual need for the quantitative assessment of colour to mimic the mechanisms of the human perceptual grasp of tones and hues, machine vision colour assessment must work on the same input information (the visible leaf) as does the human, and must conform to the same classification system and standards in its output. It is perhaps natural, then, and especially so considering the tendency of human beings to create systems after their own design (e.g. robot arms), that the digitisation and processing of images should borrow from what is known of the human visual system.

A human eye gathers light through the pupil and employs the lens under the control of the ciliary muscles to focus the light so as to produce images of objects, near or

far, on the retina. The approximately 10^8 optical sensors in the retina comprise 90% *rods*, which are sensitive to light intensities even in quite poor illumination but which have low spatial resolution and no colour sensitivity, and 10% *cones*, which occupy a limited region of the retina known as the *fovea* and which achieve very high resolution, given suitable light conditions. Figure 3.3 gives a diagram of the cross-section of a human eye, showing its relevant components.

A machine vision system must also focus the light that is reflected or emitted from objects of interest, so as to form an image on a light-sensitive surface. This surface will be an array of light-sensitive elements which are either electronic, as in the case of the CCD, or photochemical, as in the case of photographic film. The sensitive elements themselves, in both cases, may either respond only to the intensity of the incident light (yielding a grayscale image), or else they may possess a selective sensitivity to different wavelengths of the incident light, so giving rise to a full-colour representation of the objects being viewed. A set of analogies between the components of the human and of the machine-based visual systems is immediately apparent.

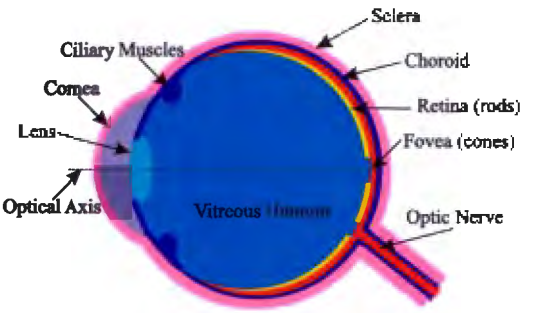


Figure 3.3: **Human eye cross section**

Research into the workings of the human eye reveals [58] that three type of cone exist, and that these are most sensitive to a certain blue, a certain green and a certain red wavelength in the visible spectrum (the "red" cones also possess some blue sensitivity, but not so much as do the "blue" cones). Figure 3.4 shows the sensitivities of each of the three cone types ($B(\lambda)$, $G(\lambda)$ and $R(\lambda)$) superimposed in a single graph of relative sensitivity against wavelength, λ. The colour perception of the human eye is then *modelled* by considering an incident ray of light, radiated or reflected from some object of interest and possessing a colour distribution $L(\lambda)$. The total responses

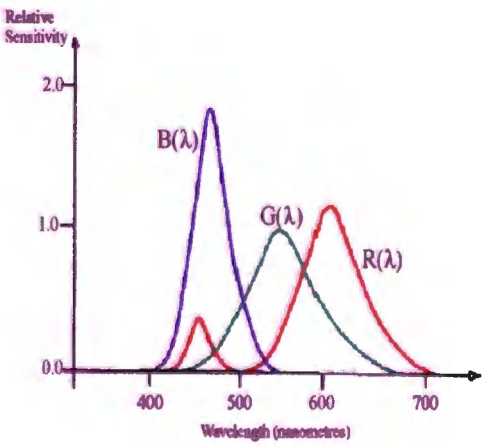


Figure 3.4: **Human eye sensitivity**

of the "blue", "green" and "red" cones to this incident ray are given, respectively, by

$$\begin{aligned} X &= C\int_{-\infty}^{\infty} B(\lambda)L(\lambda)\,d\lambda \\ Y &= C\int_{-\infty}^{\infty} G(\lambda)L(\lambda)\,d\lambda \\ Z &= C\int_{-\infty}^{\infty} R(\lambda)L(\lambda)\,d\lambda \end{aligned} \tag{3.1}$$

where C is a constant that depends upon the overall brightness response of the eye [97, 58]. The quantities X, Y and Z are known as the *tristimulus values* of the colour of the incident ray, and their sum defines the ray's total *luminance* or *intensity* that is perceived by the eye:

$$L = X + Y + Z \tag{3.2}$$

It is often useful to express the proportion of the total luminance that is represented by each of the three cone-type responses, and such proportions are called the *trichromatic coefficients* of the incident ray's colour, which are given simply by:

$$x = \frac{X}{L} \qquad y = \frac{Y}{L} \qquad z = \frac{Z}{L} \tag{3.3}$$

In specifying a colour in a machine vision system, one is faced with the problem that the human visual system itself is not completely understood, and may even differ slightly between individuals. Various models exist for the specification of the full range of perceivable colours, and each of these more or less closely represents the way that humans see colours. What almost all of the models share is that they define colours in terms of three components, which generate in a notional 3-space a region which all visible colours occupy. A particular colour may then be specified in terms of these components, and may be thought of as a point in the three-dimensional space that is spanned by the component axes. Mathematical transformations may be used to convert this specification from one colour space to another, if required. Despite this, the interpretation of the colour components that define a space will always be, at some level, a matter of definition; and this must be understood as limiting the precision of what is meant by a particular colour in a particular model. It is a fundamental assumption in this dissertation that the specification of colour, in the acquiring, digitising and processing of the images that were used, was sufficiently repeatable as to ensure that, for instance, a given pixel in a given leaf image would always be assigned the same quantised R, G and B values.

Throughout the work associated with this project, colours have been thought of in terms of a simple *RGB model*, in which each pixel in the image in considered to consist of three sensitive areas, preceded respectively by well-defined filters that pass "red", "green" and "blue" light only. The responses of the three sensitive areas are designated R, G and B, and three quantised values in the range 0-255 may then be stored to represent the pixel's colour. These values are determined on a linear scale where 0 indicates no colour content whatsoever of that particular colour, and 255 indicates maximum colour intensity for that colour component. This scheme permits the encoding of the resulting colour specification into three 8-bit words, and is thus familiarly known as *24-bit colour*. The similarity of the three encoded values to the trichromatic coefficients in the model of human vision will be obvious, and, as a matter of definition, the tristimulus values are related to the unquantised RGB values of a 24-bit colour image by the transformation [97]:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.607 & 0.174 & 0.201 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.117 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{3.4}$$

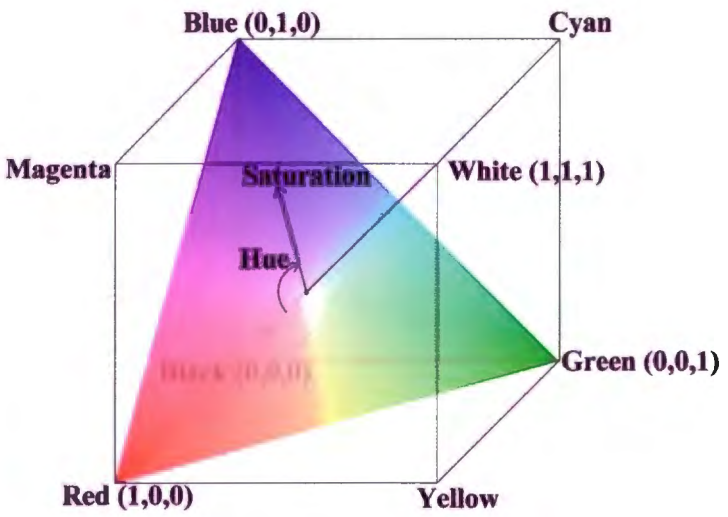


Figure 3.5: **The RGB colour space showing a plane of equal intensity in the HSI model**

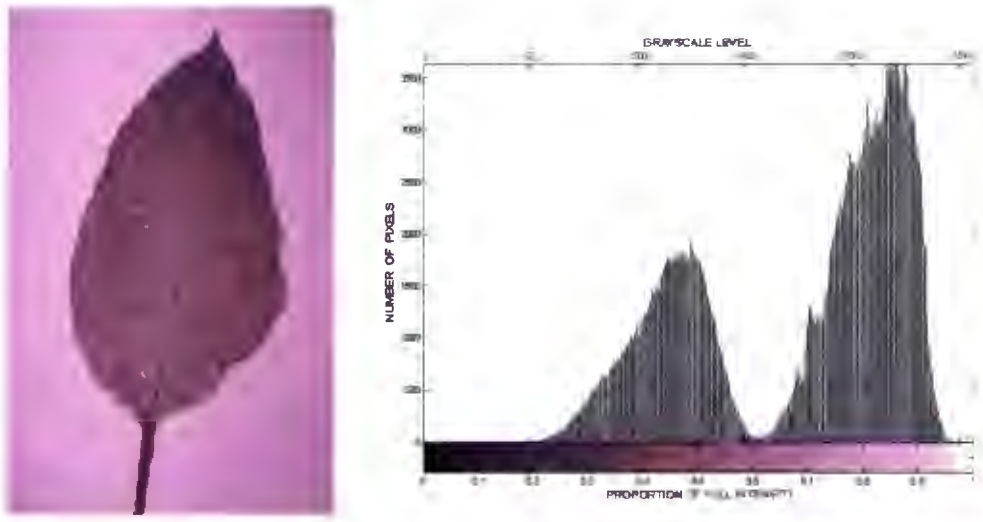
One other common colour model, which has also been of value in this project, is the *HSI colour model*, in which a colour is specified in terms of its *hue*, *saturation* and *intensity*. If one visualises the full range of specifiable colours in terms of orthogonal R, G and B components which have been normalised to their maximum possible values, then each colour would occupy a point in the unit cube shown in figure 3.5. The origin in this space represents the absence of all three colours (i.e. black), while the point $(1,1,1)$ denotes the addition of all three primary components at maximum intensity (i.e. white). In the HSI model, the diagonal line which joins black to white via every shade of grey is viewed as the *intensity* axis. If a certain colour is represented by a point in HSI space, its *saturation* is then given by the perpendicular distance of the point from the intensity axis, while its *hue* is specified by the angle of the point, measured clockwise from the line joining the central point to the red vertex as viewed from the origin. A locus of points of equal intensity and equal saturation is known as a *hue circle*. Figure 3.5 illustrates these concepts by showing a plane of equal intensity and specifying the saturation and hue of a colour represented by a point in the plane.

One advantage of working with an HSI model of colour is that it seems to follow the human visual system in giving separate treatment to luminance (intensity) and chrominance (hue and saturation). The model also allows close control to be kept on the hue component during image acquisition or processing, which is especially necessary if, as in this project, accurate colour representation is important. A minor disadvantage of the HSI scheme is that it is not as commonly used nor as simple to visualise as the RGB model. Transformation formulæ between all the components of the HSI and RGB models exist, but the only one used in this dissertation defines intensity as:

$$I = \frac{R+G+B}{3} \tag{3.5}$$

3.3 Isolation of objects within images

Once an image has been given a digital representation that has been stored in a computer, it is an easy matter to conduct statistical surveys of its pixels. In the case of the grayscale image of a leaf shown in figure 3.6(a), for example, each pixel possesses an intensity level in the range 0-255. A graph that shows the total number of pixels of each intensity level surveyed over the whole image is known as a *histogram*, and the



(a) Grayscale image (b) Histogram of the grayscale image

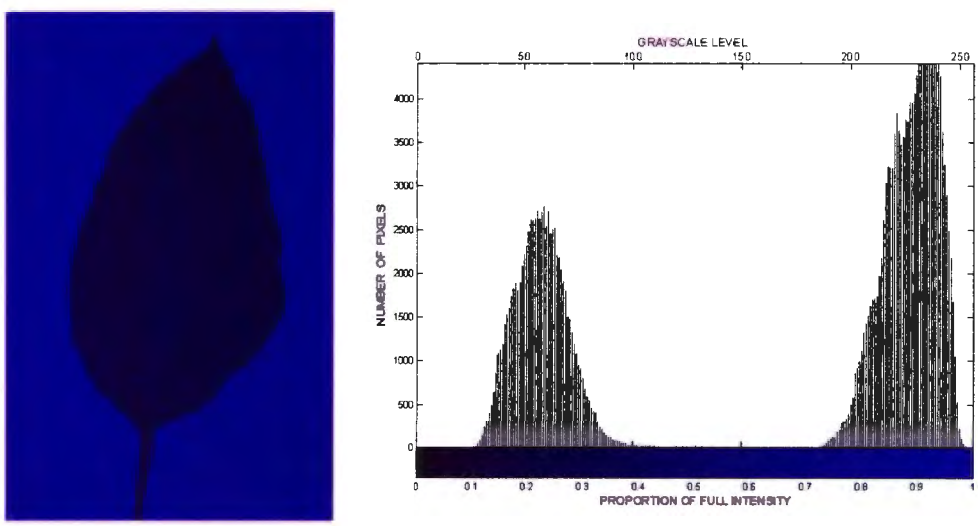
Figure 3.6: **A grayscale image, and its histogram**

histogram for the image in figure 3.6(a) is given in figure 3.6(b). The intensity histogram of this image clearly manifests a bimodal distribution, peaking firstly near an intensity level of 122 and secondly in the region around intensity level 219. These two peaks correspond to the leaf-area *foreground* and to the light-coloured *background* of the image respectively.

It is a common aim in digital image processing to wish to separate a foreground *object* from its surrounding background for the pupose of further analysing or processing the pixels in the object alone. This isolation of an object of interest is sometimes called *segmentation*. A very frequently used method of segmentation (and the exclusive method used for isolating leaves from background in this project) is to determine, for each pixel in the image, whether it contributes in the contiguous spatial sense to an object or to the background by measuring upon which side of some criterion grayscale value its intensity falls. This is a simple and powerful technique known as *thresholding*. In more formal terms, thresholding an image $f(x,y)$ at some threshold intensity criterion K (where $0 < K < 255$) returns a new image $f'(x,y)$ in which the pixels are given one of only two distinct graylevels, g_1 or g_2, such that:

$$f'(x,y) = \begin{cases} g_1 & : \quad f(x,y) \leq K \\ g_2 & : \quad f(x,y) > K \end{cases} \tag{3.6}$$

Thresholding an image at a well-chosen grayscale criterion value thus seems to mimic the power that the human brain-eye system, with its rapid grasp of intensity variation, colour difference and pictorial context, possesses for distinguishing objects from their background. Even so, thresholding is far from infallible. It may be noted that, in figure 3.6(b), there is no value of pixel intensity in the range 122-219 that does not occur somewhere in the image, and that the choice of threshold criterion, K, must be made with some care if segmentation is to be effective. If too high a value is chosen for K, then some pixels from the background will be misclassified as contributing to the object; and too low a value of K will lead to pixels that truly belong to the object being misclassified as background. Effective methods for optimising the choice of K do exist, and these include choosing the intensity value corresponding to the minimum point of the histogram curve between the two peaks, or choosing the intensity value that minimises the estimated error by using *a priori* knowledge of the probabilities that a pixel belongs to the object or to the background.



(a) Blue band image

(b) Histogram of the blue band image

Figure 3.7: **A blue band image, and its histogram**

In this project, such techniques were not found to be necessary, because images of leaves were acquired under well-controlled lighting conditions, with each leaf being set against a well-contrasted background. Nevertheless, some improvement in the separation of the two histogram peaks was observed if, instead of calculating the histogram of grayscale intensities as described above, a histogram of the blue (B) component values of every pixel in the image was considered (see figure 3.7). This

improvement was expected as a consequence of the fact that cured tobacco leaves are yellowish to brownish in overall colour, and appear in images with pixel values that are therefore low in the complementary blue component band. The contrast between the pixels representing the leaf and those of the high intensity (i.e. high *R*, *G and B*) background is thus enhanced by working with the blue band histogram, which in turn makes the choice of a threshold criterion level for the unambiguous segmentation of the leaf from the background somewhat easier.

(a) Grayscale inverted mask (b) Blue band mask

Figure 3.8: **Masks of the leaf image**

It is common to normalise the pixel values in a thresholded image by dividing throughout by the larger of g_1 and g_2 so as to obtain a *binary thresholded image* consisting only of zeros and ones with, for example, the zeros all corresponding to the background in the original image, and the ones in the positions formerly occupied by the object. Such an array is known as a *mask*, and figure 3.8(a) illustrates the mask that was derived by thresholding the image in figure 3.7(a) using a grayscale criterion value of 161, while figure 3.8(b) shows the result when thresholding was done using the same intensity criterion of 161 in the blue (*B*) band. One can see that the results are very similar, but that the small amount of "salt-and-pepper" noise due to incorrect separation of object from background in figure 3.8(a) has been effectively removed in figure 3.8(b) through the use of the superior thresholding method. Figure 3.8(a)

uses white to represent zeros and black for ones, and could therefore be viewed as an *inverted mask.*

The value of creating a mask that corresponds to an object in an image in this way is that the mask can then be used as a logical operand with which to isolate the pixels of the object of interest for the purpose of further analysis, processing or display. Thus, for instance, by treating the white pixels in figure 3.8(b) as ones and the black pixels as zeros and then performing a logical AND of every pixel in the mask with the identically-positioned pixel in the image of figure 3.9(a), a new image consisting only of the object of interest (the leaf) can be derived as shown in figure 3.9(b). This completes the isolation (or segmentation) of the object from the rest of the image.

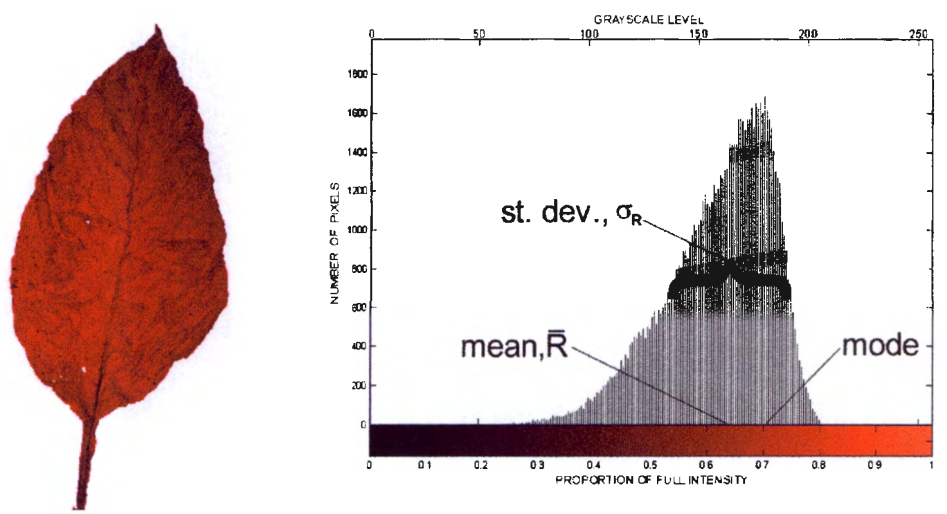
(a) The object plus background (b) The segmented object

Figure 3.9: **The use of a mask to segment an object from its background**

3.4 Characterising the colour of an object

It is possible to survey the pixels of an isolated object, much as was done for an entire image in the previous section. This can lead to the extraction of statistical parameters which give a good specification of the object's colour. In the case of the leaf object

in figure 3.9, a simplistic overall assessment of its colour as "brownish" or "yellowy-brown" will certainly not yield the quantitative evaluation of colour that is required for repeatable machine vision grading. Furthermore, a simple averaging of the R, G and B values of every pixel in the object, while providing a quantitative measure of overall colour ($\overline{R}$, $\overline{G}$, $\overline{B}$), will not express the degree to which the colour has been darkened by damage to certain areas of the lamina during the curing process. In fact, no single-valued measure of the leaf's colour can fully convey the variation in the hue and tone of the leaf across its surface. If a machine vision characterisation of the colour of the leaf is to have any chance of competing with the human visual system's almost immediate appreciation of both integrated overall colour and differentiated colour variation within the leaf object, then it will have to work with a number of parameters, preferably derived from a survey of *all* of the pixels in the object.



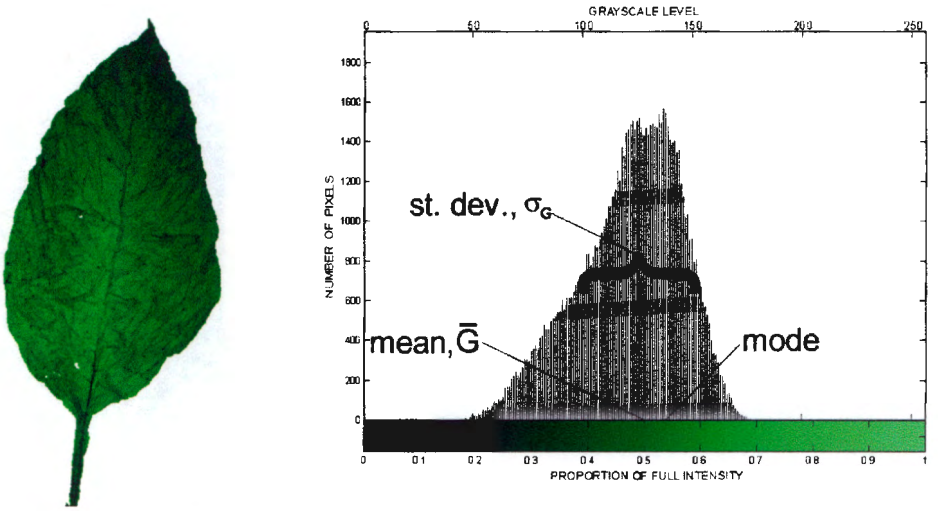
(a) Object in red band

(b) The object's red band histogram

Figure 3.10: **The leaf object in the red band, and its histogram**

Parameters of this nature are readily extractable from histograms of the colour component bands compiled from the pixels that make up the object. So, for example, rather than working with only the *mean* red value, $\overline{R}$, of the leaf object's pixels, the histogram of the red (R) band (see figure 3.10) also yields parameters such as the *modal* red value, R_{md} (indicative of the red content of the most commonly observed lamina colour of the leaf), and the red value *variance*, σ_R^2 (which suggests the degree of red content variation in pixels throughout the leaf, perhaps due to partial lamina damage). These parameters are illustrated in figure 3.10(b), where the *standard deviation*, σ_R,

which is simply the square root of the variance, is given in order to present a measurement that is in the same units as the histogram data. In a similar way, the histogram of the green (G) band may be used to derive mean ($\overline{G}$), modal (G_{md}) and variance (σ_G^2) values that assist in parameterising the colour of the leaf. Figure 3.11 illustrates this for the leaf object of figure 3.9. Corresponding values extracted from a histogram



(a) Object in green band

(b) The object's green band histogram

Figure 3.11: **The leaf object in the green band, and its histogram**

of the blue (B) band were not used in this project, but the mean ($\bar{I}$), mode (I_{md}) and standard deviation (σ_I) of the histogram of pixel intensities (see figure 3.12) did prove to be of value, as described in a later chapter. It will be seen from equation 3.5 that this is tantamount to using the information contained in the blue (B) band.

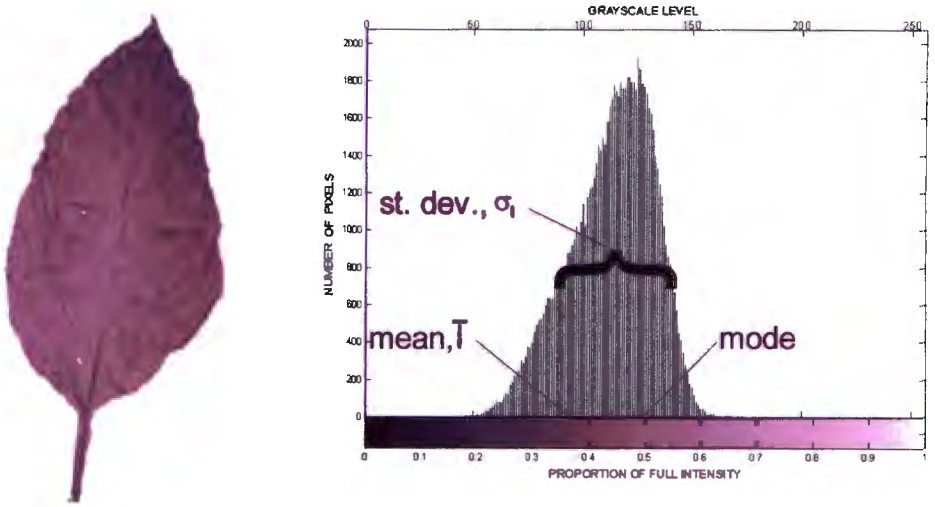
In all of these colour parameters, using x to denote R, G or I, the *mean* is calculated as

$$\bar{x} = \frac{1}{N} \sum_{i=0}^{255} n_i i \tag{3.7}$$

where

$$N = \sum_{i=0}^{255} n_i \tag{3.8}$$

is the total number of pixels in the object, and n_i denotes the number of pixels within the object that have component band value i. The *mode* is the value of i within the



(a) Object intensity (b) The object's intensity band histogram

Figure 3.12: **The leaf object in the intensity band, and its histogram**

distribution of x over i for which n_i is a maximum for $0 \leq i \leq 255$. The *variance* is then the well-known measure of distribution spread:

$$\sigma_x^2 = \frac{1}{N} \sum_{i=0}^{255} n_i (i - \bar{x})^2 \tag{3.9}$$

The representation of an object's pixels by a histogram offers a technique called *histogram equalisation* which is very useful for enhancing the *contrast* (i.e. the mean grayscale value difference) between the pixels. This technique may make features that were almost invisible in the un-equalised object into clearly distinguishable spatial distributions of pixels which may then, themselves, be segmented from their surroundings (perhaps by thresholding) and analysed as objects in their own right. Histogram equalisation aims to transform the histogram of an object, such as the intensity band object of figure 3.12, into a histogram in which each grayscale value in the range 0-255 is equally likely to occur. Because of the quantised nature of the intensity distribution in the un-equalised histogram, however, a truly uniform distribution in the equalised histogram cannot be achieved by a simple mapping function. Nevertheless, substantial enhancement of the image may be achieved by mapping the grayscale value $g_{(x,y)}$ of every pixel in the un-equalised image $f(x,y)$ to a new grayscale value $g'_{(x,y)}$ in the

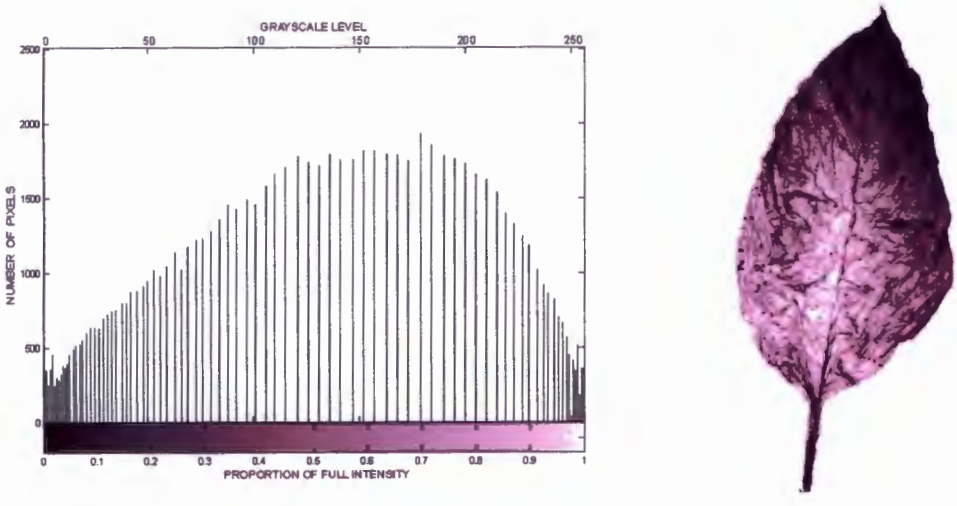
equalised image $f'(x,y)$ using the mapping function

$$g'_{(x,y)} = 255 \sum_{i=0}^{g(x,y)} G_i \tag{3.10}$$

where G_i is the proportion of pixels in the un-equalised image that are of grayscale value i, given by

$$G_i = \frac{n_i}{N} \tag{3.11}$$

Figure 3.13 shows the equalised histogram resulting from using this mapping on the histogram of figure 3.12, and also illustrates the corresponding leaf object, in which the veinous system of the leaf is now considerably enhanced, and may be considered for further processing or segmentation as described later.



(a) The equalised histogram (b) Equalised leaf

Figure 3.13: **Histogram equalisation, and its effect in enhancing the image of a leaf**

3.5 Characterising the size and shape of an object

Perhaps the simplest size characteristic to extract from a single contiguous segmented object is its area — for this can be stated merely as the number of pixels that the object comprises. This number could be arrived at, for example, by taking the histogram of

the binary thresholded image in which the object's pixels have been set to a grayscale value of 255, and then calculating n_{255}. An area so calculated remains in units of *pixels*, unless a conversion factor to conventional units of area exists, either through a known resolution (e.g. in pixels per inch) for the image once printed or, as in figure 3.14, if a scale or object of known absolute dimensions has been included in the image for the express purpose of calculating such a conversion. Thus, figure 3.14 is printed on page 47 with a resolution of 200 pixels per inch, so that the $N = 86693$ pixels in the leaf object are represented by an area of

$$\left(\frac{86693}{200^2}\right) = 2.167 \text{ [in}^2\text{]}$$

or 13.98 [cm]^2 on the printed page.

Alternatively, the solid black block which was photographed in the corner of the image is a square of side exactly 2 cm. When this image was first scanned it had a resolution of 1700 pixels by 2600 pixels, and the square block appeared in the image with sides of length 75 pixels. Subsequent resampling of the image has reduced it to pixel dimensions of 400 by 612 (i.e. by a factor of 4.25), so the most accurate calculation of the area of the black square would now be

Figure 3.14: **Scaled image**

$$\frac{75^2}{4.25^2} = 311.419 \text{ [pixels]}$$

Hence, 1 cm^2 comprises 311.419/(2^2)=77.855 pixels, and so the $N = 86693$ pixels in the leaf object represent a true area of the original leaf of 86693/77.855=1113.5 cm^2. From this, the linear *scale* of the leaf diagram in figure 3.14 may be calculated as

$$\sqrt{\frac{1113.5}{13.98}} \; : \; 1 \; = \; 8.92 : 1$$

The *length* and *width* of an object parallel to the image axes are similarly simple to derive once the object has been segmented. In principle, an inspection of the coordinates (x, y) of each pixel in the object referred to an origin at the corner (say the top left-hand corner) of the whole image will reveal the minimum and maximum values

of x and y that are found anywhere within the object. Then the vertical length l and horizontal width w of the object are given by

$$l = y_{max} - y_{min} \text{ [pixels] and } w = x_{max} - x_{min} \text{ [pixels]} \tag{3.12}$$

and these distances can now easily be converted to centimetres, for example using the conversion factor of $\sqrt{77.855} = 8.824$ pixels per centimetre. Furthermore, because the pixels of the image are presented in a cartesian (x,y) array, the distance between *any* two pixel points $A(x_A,y_A)$ and $B(x_B,y_B)$ can be found very simply from the Euclidean distance metric:

$$\overline{AB} = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2} \text{ [pixels]} \tag{3.13}$$

Since the pixels whose co-ordinates contain the four extreme values of x and y will inevitably lie on the outer perimeter of the object, these extreme pixels could be found (although this would not be the fastest method to use) through the use of a *boundary finding algorithm.* Such an algorithm might begin at the *centroid* $(\bar{x},\bar{y})$ of the object, defined in terms of the number of pixels N in the entire object as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.14}$$

The algorithm then increments the value of one of the co-ordinates (say the x co-ordinate) until it reaches a point in the binary thresholded image at which a transition occurs from a pixel intensity value representing the object to a value representing background. This point lies on the boundary of the object. All such points may be identified by means of a number of existing *edge detection* algorithms, which operate on binary, grayscale or colour images as two-dimensional digital differentiators of intensity as a function of the co-ordinate variables, x and y. A suitable and simple detector for the purpose of isolating the boundary of a binary (0,1) thresholded image employs the *filter mask* shown below:

-1	-1	-1
-1	8	-1
-1	-1	-1

This mask is passed over every pixel point in the image whose object boundaries are to be detected. At each point (x_c,y_c) in the binary thresholded image $f(x,y)$, the filter

returns the value $g(x_c,y_c)$, where

$$\begin{aligned} g(x_c,y_c) \;=\; & -f(x_c-1,y_c+1)-f(x_c,y_c+1)-f(x_c+1,y_c+1) \\ & -f(x_c-1,y_c) \quad +8f(x_c,y_c) \quad -f(x_c+1,y_c) \\ & -f(x_c-1,y_c-1)-f(x_c,y_c-1)-f(x_c+1,y_c-1) \end{aligned} \tag{3.15}$$

as suggested by the filter weights in the filter mask. Subsequent thresholding of $g(x,y)$ can then be used to return the edge-detected image to binary (0,1) form. Figure 3.15 shows the result of performing this filter operation on the binary (0,1) thresholded image of figure 3.8(b). It will be noted that this algorithm has found the object's edge to a positional accuracy of one pixel.

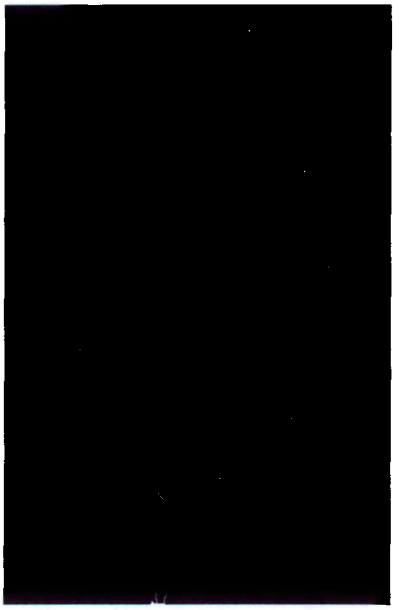


Figure 3.15: **Detected object outline**

Once the boundary has been identified in this way and then segmented as an object, only those pixels that are in the boundary object need to be considered in order to find the extrema of x and y within the original object. Moreover, the *perimeter* of the object may now be estimated in several ways. One method of perimeter measurement considers the absolute length of the object exterior by tracing a path around all of the pixels in the boundary, adding the exterior pixel lengths until returning to the initial point. A quicker and surprisingly accurate alternative is to take the total number of pixels in the boundary and then treat that number as the perimeter length. This is justified so long as the boundary is no more than one pixel thick (as would be the

case with the thresholded output of the filter method described above), and assuming adjacent boundary pixels to neighbour one another in the 4-neighbours sense (i.e. vertically or horizontally, again as in the method above, but not diagonally). The absolute object exterior in figure 3.15 was estimated at 1794 pixels by a boundary tracking algorithm, and the total number of pixels in the boundary is 1793. This happens to be an excellent agreement — in general, the two methods were found by experience to agree within 1%. These estimates mean that the perimeter of the leaf in figure 3.14 has a length of 1794/200*2.54=22.8 cm in the printed image and $1794/8.824 = 22.8 \times 8.92 = 203.3$ cm in the actual leaf.

Further geometrically or statistically derived quantities may be extracted from basic image length measurements in very natural ways. So, for instance, given the points A, B and C in an image, the angle $\widehat{BAC}$ is found, as might be expected, from the Cosine Rule:

$$\cos\widehat{BAC} = \frac{\overline{AB}^2 + \overline{AC}^2 - \overline{BC}^2}{2\,\overline{AB}\,\overline{AC}} \tag{3.16}$$

Similarly, having measured a sequence of widths of an irregular object, for example, characterisation of the shape of the object by statistics such as the mean or the variance of these widths is a natural step. Both of these types of derived quantities were used to advantage in this project, as discussed in later chapters.

3.6 Morphological filters

Morphological image processing provides a class of non-linear filters that can be used to operate on objects within images, altering their geometrical structure. The simplest morphological techniques are intended for use with binary images only, but grayscale morphological procedures also exist, and, since both binary and grayscale morphological filters were used in this project, they will both be described here in some detail.

Binary *dilation* is a process which augments the number of pixels that represent an object within an image, with each pixel in the original un-dilated object being replaced in the dilated object by several pixels, as determined by a *structuring element.* The structuring element is thought of as operating upon the pixels of the image, replacing each pixel with a map of its own pixel arrangement, after which the dilated output image is taken to be the union of all of the maps so formed. As with all such filters,

the action of dilation is most elegantly explained in terms of *set theory*. Thus, if an object $\mathcal{A}$ is dilated by a structuring element $\mathcal{B}$, both $\mathcal{A}$ and $\mathcal{B}$ may be treated as sets of *elements*, denoted individually by a and b.

One way of viewing dilation is to regard each element of set $\mathcal{A}$ as being translated through a series of vector shifts from the origin of $\mathcal{A}$, as indicated by the positions of the nonzero pixels in set $\mathcal{B}$ with respect to the origin of $\mathcal{B}$. The dilated version of $\mathcal{A}$ is then written as $\mathcal{A} \oplus \mathcal{B}$, and is taken to be the union of all the translated versions of $\mathcal{A}$. With t representing a translated set occupying the two-dimensional space $\mathbb{I}^2$, the set theoretic expression for dilation is then

$$\mathcal{A} \oplus \mathcal{B} = \bigcup_{b \in \mathcal{B}} \{t \in \mathbb{I}^2 : t = a + b,\ a \in \mathcal{A}\} \tag{3.17}$$

where "+" represents the vector addition of the offsets of nonzero elements in $\mathcal{B}$ to the nonzero elements of $\mathcal{A}$.

As an example, the object illustrated below, when repeatedly offset by the shifts indicated by the nonzero pixels of the structuring element shown (in which the asterisk represents the origin), yields the dilated image on the right hand side.

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	0	1	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

$\oplus$

*0	1
1	1

$=$

0	0	0	0	0	0	0
0	0	1	0	0	0	1
0	1	1	1	0	1	1
0	0	1	1	1	1	0
0	0	0	1	1	1	0
0	0	1	1	1	1	1
0	1	1	1	1	1	1

It may be noted that this particular structuring element happens to yield an output image whose centre is shifted one pixel down and one to the right in the frame of the object. In larger objects this effect would probably be negligible. The second notable point is that exactly the same dilated image output can be arrived at by thinking of the structuring element, rather than the original object, as being the "mobile" operand. If structuring element $\mathcal{B}$ is submitted to the co-ordinate mapping

$$\{x, y : x \to -x,\ y \to -y\}, \tag{3.18}$$

then it will appear as shown here, and may be written $-\mathcal{B}$:

1	1
1	*0

Now applying $-\mathcal{B}$ to $\mathcal{A}$ as a travelling mask (much as was described for boundary finding in the previous section), dilation is achieved by taking the union of all points that yield a nonzero result (i.e. some overlap) for the intersection of $\mathcal{A}$ and $-\mathcal{B}$ when the asterisked origin of $-\mathcal{B}$ lies above them. Again, set theory gives concise expression to this as

$$\mathcal{A} \oplus \mathcal{B} = \bigcup_{x \in \mathbb{I}^2} \{\mathcal{B}_x \cap \mathcal{A} \neq \emptyset\} \tag{3.19}$$

where

$$\mathcal{B}_x = \{t \in \mathbb{I}^2 \,:\, t = b + x,\ b \in -\mathcal{B}\} \tag{3.20}$$

Binary *erosion* differs from dilation in that it uses a structuring element to *reduce* the number of pixels in an object. One can visualise moving the object $\mathcal{A}$ as indicated by the nonzero pixel positions in $-\mathcal{B}$ so as to create a series of maps of $\mathcal{A}$. The eroded image of $\mathcal{A}$ will then consist only of those pixels which are nonzero in *all* of the maps — in other words, the eroded image is the intersection of the translated versions of set $\mathcal{A}$. This view of the erosion of $\mathcal{A}$ by $-\mathcal{B}$ is summarised as

$$\mathcal{A} \ominus \mathcal{B} = \bigcap_{b \in -\mathcal{B}} \{t \in \mathbb{I}^2 \,:\, t = a + b,\ a \in \mathcal{A}\} \tag{3.21}$$

and is illustrated below for the same object $\mathcal{A}$ and structuring element $-\mathcal{B}$, using the symbols ı and ıı to indicate where only one or two respectively of the three maps gave a nonzero result for a location.

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	0	1	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

$\ominus$

1	1
1	*0

$=$

ı	ı	0	0	ı	ı	0
ı	ı	ı	ı	ıı	0	0
0	ı	ı	ıı	0	0	0
0	ı	1	ıı	ı	0	0
ı	1	1	1	ıı	ı	0
ı	ı	ı	ı	ı	0	0
0	0	0	0	0	0	0

Careful inspection of the object $\mathcal{A}$ and non-inverted structuring element $\mathcal{B}$ reveals that exactly the same eroded result is achieved by using $\mathcal{B}$ as a mobile mask and identifying

all positions in $\mathcal{A}$ for which every nonzero element of $\mathcal{B}$ sits above a nonzero element of $\mathcal{A}$. This completes a pleasing symmetry by adding the result:

$$\mathcal{A} \ominus \mathcal{B} = \bigcap_{x \in \mathbb{I}^2} \{\mathcal{B}_x \cup \mathcal{A} \neq \emptyset\} \tag{3.22}$$

where

$$\mathcal{B}_x = \left\{t \in \mathbb{I}^2 \,:\, t = b + x,\, b \in \mathcal{B}\right\} \tag{3.23}$$

Erosion, like dilation, tends (with an asymmetric structuring element) to shift the output image very slightly, as the example above illustrates. Both erosion and dilation can also have a dramatic effect on the area of the image upon which they operate, but again this effect may be negligible if the structuring element is much smaller than the object being eroded or dilated. Figure 3.16 shows how the use of a structuring element in the form of a circular disk of radius 9 pixels acts when dilating the image of a leaf. Because the structuring element is symmetrical, there is no shifting of the output image, and because it is small, the effect on the output image's area is minimal. It is interesting to note that the outline of the output image is the locus of points that would have lain on the outer perimeter of the circular structuring element as it was rolled along the input image, with its centre following the outer contour of all the objects (even noise). The operation tends to blur the finer detail of the leaf outline.

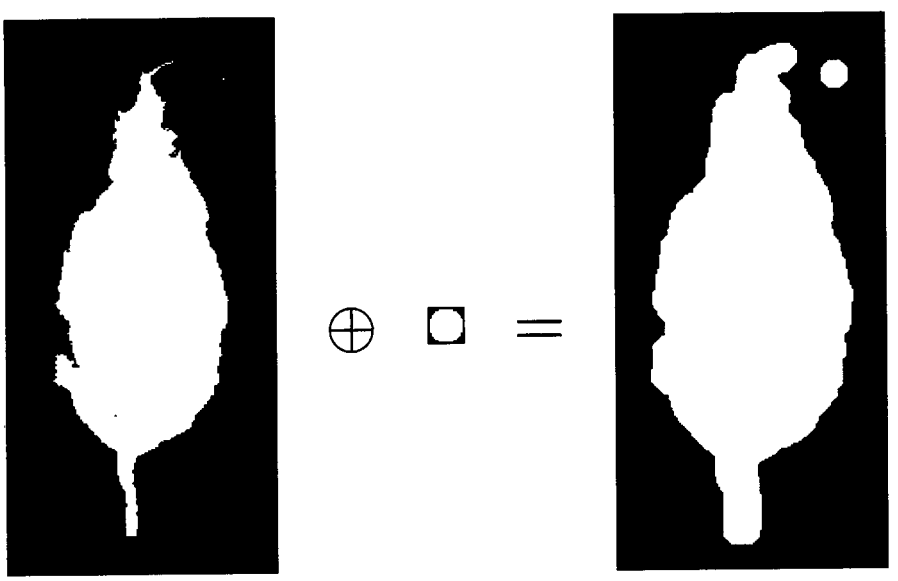


Figure 3.16: **Dilation of a leaf image by a circular disk of radius 9 pixels**

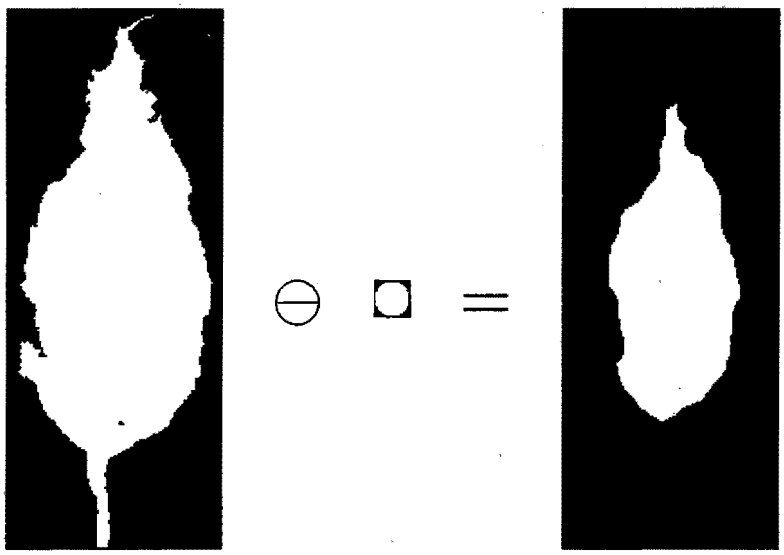


Figure 3.17: **Erosion of a leaf image by a circular disk of radius 9 pixels**

In figure 3.17, the same leaf is shown being eroded by the circular disk. The output image is, in effect, the remainder of the leaf after the centre of the structuring element has been rolled along the outer leaf contour and the structuring element has "swept away" the parts of the leaf over which it has passed. It will be noted that the output object is smaller than the original leaf, with its exterior a smoothed copy of the leaf contour. Thin protruding boundary features, and notably the butt of the leaf, have been removed by the erosion.

Further morphological operations have been developed which preserve the size and position of the objects upon which they act, whilst retaining some of the useful effects that erosion and dilation have on the object contour. One such procedure is morphological *opening*, which is, quite simply, erosion followed by dilation, usually with the same structuring element. Hence, if $\mathcal{A}$ is opened with $\mathcal{B}$, the result is written

$$\mathcal{A} \circ \mathcal{B} = (\mathcal{A} \ominus \mathcal{B}) \oplus \mathcal{B} \tag{3.24}$$

where the brackets indicate the order in which the operations must be carried out. This is performed below on the small image of the previous example, and it can be seen that the opening has smoothed the outer contour of the input by removing its thin external features whilst broadly restoring the size and position of the object.

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	0	1	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

$\circ$

*0	1
1	1

$=$

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	0	0
0	0	0	0	0	0	0

The effect on the object outline is easier to observe in a bigger image, such as is shown in figure 3.18, where the circular structuring element of radius 9 pixels has now been used to open the outline of a tobacco leaf. This has somewhat smoothed the boundary of the object and has removed the protruding features (including the butt of the leaf) whilst keeping the leaf's size and position as they were.

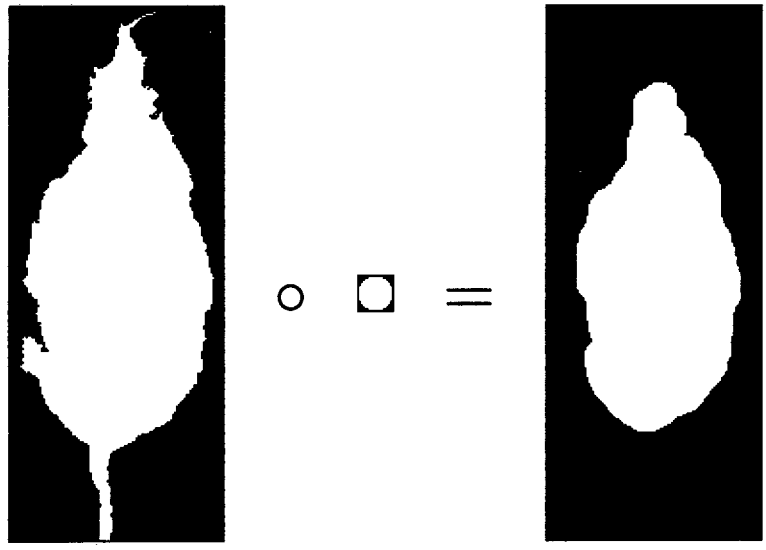


Figure 3.18: **Opening of a leaf image by a circular disk of radius 9 pixels**

When dilation is followed by erosion, the resulting combined operation is known as morphological *closing* and is written as

$$\mathcal{A} \bullet \mathcal{B} = (\mathcal{A} \oplus \mathcal{B}) \ominus \mathcal{B} \tag{3.25}$$

The equation below shows the result when the same structuring element $\mathcal{B}$ is used to close the example object $\mathcal{A}$. What is seen here is that closing has retained the thin exterior features of the object, whilst "filling in" on the right hand side the narrow bay in the object's outline contour. Size and position are, again, little affected.

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	0	1	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

$\bullet$

*0	1
1	1

$=$

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	1	0	1	0	0
0	0	0	1	1	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

Closing an image will generally smooth the outline by filling small concavities, as illustrated in figure 3.19, where the same structuring disk as before has been employed to close a leaf image. It will be observed how the exterior features of the leaf have been retained, except for the closure of narrow bays in the outline.

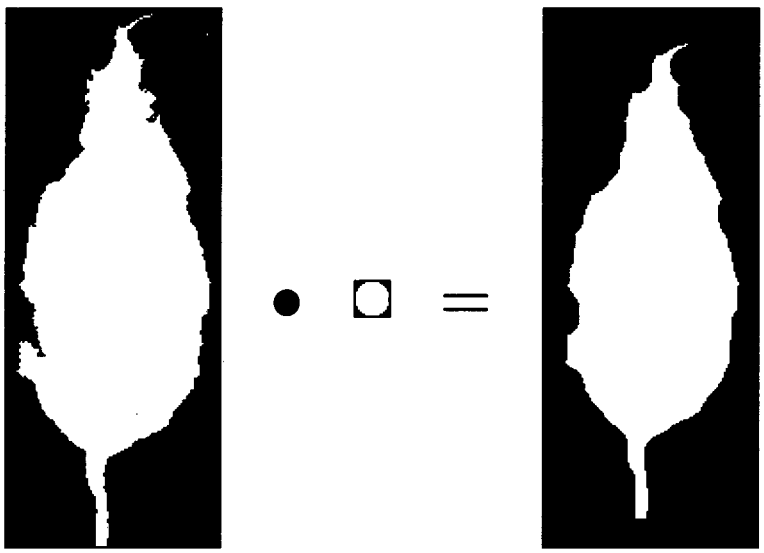


Figure 3.19: **Closing of a leaf image by a circular disk of radius 9 pixels**

The morphological operations discussed above are defined solely as processing techniques for use on binary images, but morphological concepts may be extended to be of value in the processing of grayscale images as well. In grayscale erosion, for example, both the image to be eroded and the structuring element may be thought of as three-dimensional surfaces whose height at any point is given by the intensity value of the pixel at that point. The structuring element is then passed as a mask over the grayscale image by placing its (asterisked) origin or *active point* over each pixel of the image in turn. In each position, every pixel value of the mask is subtracted from the

corresponding pixel value of the image which it overlies, and the *minimum* of all the subtracted values is returned as the pixel intensity value at the corresponding active point in the eroded output image. Mathematically, each element of the eroded output image, $g(x,y) = f(x,y) \ominus s(x,y)$, is found at position (x_c, y_c) from the input image $f(x,y)$ and $m \times n$ stucturing element $s(x,y)$ as

$$g(x_c,y_c) = \min_{i=0,\ j=0}^{i=m,\ j=n} \{f(x_c+i,y_c+j) - s(x_c+i,y_c+j)\} \tag{3.26}$$

Likewise, dilation of a grayscale image is achieved by returning to the active point of the output image the *maximum* value after addition of each of the structuring element pixel values to the corresponding image pixel value, for every position of the active point of the mask. In this case, each element of $g(x,y) = f(x,y) \oplus s(x,y)$ is given by

$$g(x_c,y_c) = \max_{i=0,\ j=0}^{i=m,\ j=n} \{f(x_c+i,y_c+j) + s(x_c+i,y_c+j)\} \tag{3.27}$$

It will be noted that in positions where the output intensity value lies outside the permissible range of 0-255, a saturated pixel value of 0 or 255, as appropriate, is returned. Edge effects at the periphery of the image are minimised by invoking a surrounding skirt of pixels or by implementing a wrap-around, as desired.

Figure 3.20(a) shows a small grayscale image, which consists of a background of intensity value 128 and two objects, similar to the one used in the earlier example, of intensity values 64 and 192. For extra realism, additive noise of rms intensity value 25 has been applied to this image, to yield the grayscale image of figure 3.20(b).

(a) A grayscale image (b) Grayscale image with noise

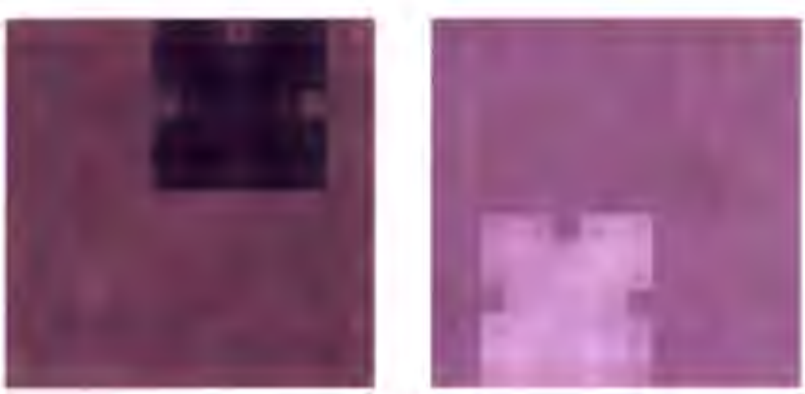
Figure 3.20: **Image used to illustrate grayscale morphology**

The structuring element

*10	20	10
20	30	20
10	20	10

was used to perform both grayscale erosion and dilation of this image, and the results are shown in figure 3.21. It is immediately clear that the erosion has darkened the image background and has removed the lighter (higher) of the two objects in the input image. This is consistent with viewing the structuring element as having been inverted (so higher numbers are "deeper") and then used to scour out the surface of the input image. This has "deepened" the depression representing the darker object, almost obliterated the lighter object, and evenly sliced the background to a depth of about 30 below its original contour. A smoothing of the surface, analogous to the smoothing of lines in binary erosion, may also be noted.

Less obviously, but of importance here, is that the structuring element has imposed some of its symmetry upon the objects in the eroded image, leaving the dark object almost a copy of the size and depth below the background of the (inverted) structuring mask. In certain cases, preferential retention of some of the features of the input image can be ensured by choosing a structuring element which mimics the object features which one wishes to preserve — a fact that was used to some advantage in this project, as reported in a later chapter.

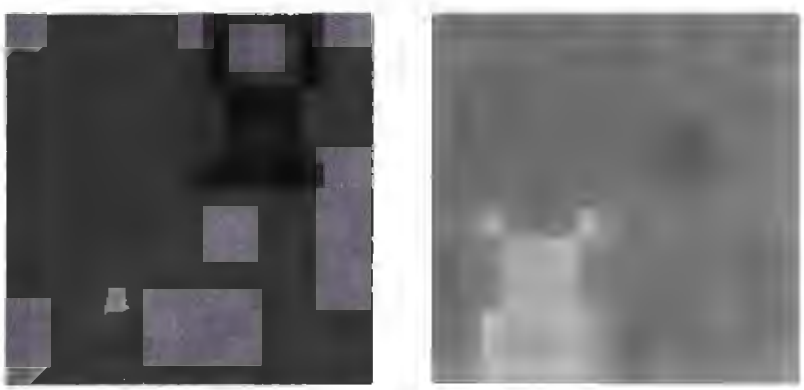


(a) After grayscale erosion (b) After grayscale dilation

Figure 3.21: **The effects of grayscale erosion and dilation**

Figure 3.21(b) shows the output of the grayscale dilation, which can be interpreted as having achieved a "filling" of the surface in three dimensions to the depth of the structuring element. Thus, the background has lightened almost uniformly by about 30 intensity points, the darker object has been almost obliterated and the lighter object intensified and made to conform to the symmetry of the mask.

Some of the intensity and shape distortion imposed by grayscale erosion and dilation is corrected in the operations of grayscale opening and closing. These are defined in the natural way as erosion followed by dilation for opening, or *vice versa* for closing, and the results for the same example object and structuring element are shown in figure 3.22. One can see that one object is present in each of the output images after opening or closing, and that the background has been restored to its previous intensity. Opening has enhanced the peripheral fine structure of the darker object, while closing has preserved the gross size of the lighter object. The precise value of these results would depend on the application in which the grayscale morphological operators were used, and also on the size of the structuring element in comparison with the input image objects. The illustration above was made at relatively low resolution — at high resolution the remaining distortion of the objects in figure 3.22 would probably be negligible for most applications.



(a) After grayscale opening (b) After grayscale closing

Figure 3.22: **The effects of grayscale opening and closing**

3.7 Geometrical transformations

Multi-stage image processing algorithms frequently include one or more subroutines whose function is to alter the geometrical structure of the image, or of objects within the image. A simple but important example of this is *translation*, in which the position of an object in an image is adjusted with respect to a notional origin without otherwise affecting the shape or size of the object and with minimal disturbance to the background. Each pixel (x_t, y_t) in the translated output image $g(x,y)$ is considered in turn, and its corresponding pixel (x_c, y_c) in the input image $f(x,y)$ is found as

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{cases} \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \begin{bmatrix} x_o \\ y_o \end{bmatrix} & : \text{ if } (x_t - x_o, y_t - y_o) \in O \text{ in } f(x,y) \\ \begin{bmatrix} x_t \\ y_t \end{bmatrix} & : \text{ otherwise, and if } (x_t, y_t) \notin O \text{ in } f(x,y) \\ \text{undefined} & : \text{ otherwise, and if } (x_t, y_t) \in O \text{ in } f(x,y) \end{cases} \tag{3.28}$$

where O is the set of pixels in the input image which comprise the object to be translated, and x_o and y_o are the desired vector translations in the $x-$ and $y-$directions respectively. Translation is then achieved by setting $g(x_t, y_t) = f(x_c, y_c)$ for all defined values of (x_c, y_c) and by returning a suitable background value (for a "copy" effect) or null pixel value (for a "cut and move" effect) in those cases where the pixel position has been vacated by the object's translation, leaving (x_c, y_c) undefined. This formulation of translation, whilst more complex than that found in most texts, is a more complete statement of the method, and will work correctly right up to the edges of the translated output image.

The technique of working backwards from the pixels of the output image to find corresponding input pixels is of particular value in encoding *rotation*. Rotation of some of the pixels in $f(x,y)$ through a clockwise angle θ about the input image origin may easily be envisaged to obtain a rotated copy of those pixels by applying

$$\begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3.29}$$

However, since in general this yields non-integer values for the pixel positions (x_r, y_r) of the output image, it is preferable to define the rotation inversely in terms of the corresponding pixels (x_c, y_c) in the input image for each output pixel position, (x_r, y_r).

Then

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{cases} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_r \\ y_r \end{bmatrix} & : \text{ if } \begin{bmatrix} x_c \\ y_c \end{bmatrix} \in O \text{ in } f(x,y) \\ \begin{bmatrix} x_r \\ y_r \end{bmatrix} & : \text{ otherwise, and if } (x_r,y_r) \notin O \text{ in } f(x,y) \\ \text{undefined} & : \text{ otherwise, and if } (x_r,y_r) \in O \text{ in } f(x,y) \end{cases} \tag{3.30}$$

and rotation is implemented, similarly to translation, by setting $g(x_r,y_r) = f(x_c,y_c)$ or to a suitable background or null pixel value if (x_c,y_c) is undefined. It will be observed that most values of θ will also yield non-integer values for x_c and y_c, so that corresponding pixel intensity values will have to be chosen either from the nearest pixel to (x_c,y_c) or, for higher quality results, by *interpolation* of the values of a set of neighbouring pixels. In this project, it has not been found necessary to use sophisticated interpolation for the required image rotations.

Rotation of an object about a point other than the origin is achieved simply by translating the entire input image so that the origin lies on the desired rotation point, performing the rotation, and then shifting the image back so as to restore the origin to its initial position. The useful operation of rotating an object through clockwise angle θ about its centroid $(\bar{x},\bar{y})$, for example, is described by the transformation

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \left(\begin{bmatrix} x_r \\ y_r \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \right) + \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \tag{3.31}$$

A third geometrical transformation is *scaling*, in which the size of an image or of an object within an image is adjusted (enlarged or reduced) by constant factors in the $x-$ and/or $y-$directions. For each pixel position (x_s,y_s) in the scaled output image, the corresponding input image pixel is

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{cases} \begin{bmatrix} \frac{1}{S_x} & 0 \\ 0 & \frac{1}{S_y} \end{bmatrix} \begin{bmatrix} x_s \\ y_s \end{bmatrix} & : \text{ if } \begin{bmatrix} x_c \\ y_c \end{bmatrix} \in O \text{ in } f(x,y) \\ \begin{bmatrix} x_s \\ y_s \end{bmatrix} & : \text{ otherwise, and if } (x_s,y_s) \notin O \text{ in } f(x,y) \\ \text{undefined} & : \text{ otherwise, and if } (x_s,y_s) \in O \text{ in } f(x,y) \end{cases}$$
$$\text{(with } S_x < 1 \text{ and/or } S_y < 1) \tag{3.32}$$

where S_x and S_y are the desired scaling factors in the $x-$ and $y-$directions respectively. In general, and especially for enlargements of objects with elaborate interior detail, scaling requires the use of interpolation techniques because of the fact that x_c and y_c will not usually be integers under this transformation. Nevertheless, the formulation given above is a powerful one: if either scaling factor is less than 1, then the object is reduced in the output image in the corresponding direction; if either scaling factor is less than 0, then the object is reflected in the corresponding axis and is enlarged or reduced in the ratio indicated by the modulus of the factor.

Translation, rotation and scaling are all *linear* transformations, and so they may be implemented in any order and are capable of analysis or description using the notation and techniques of linear algebra. One image processing method which exploits this fact, and which was of particular value in the preprocessing of images in this project, is the *Hotelling* or *Karhunen-Loève transform.* This transform provides a means for automatically rotating an object's major axis so as to align it in a preferred direction, such as parallel to one of the image's co-ordinate axes. The value of this, for example in aligning each of the objects in a series of images to ensure similar processing and comparable measurements, will be discussed in a later chapter.

The Hotelling transform begins by deriving a statistical metric known as a *covariance matrix* from the pixel positions (x_i, y_i) of the object to be aligned. This matrix may be written as

$$\mathbf{C} = \begin{bmatrix} C_1 & C_2 \\ C_3 & C_4 \end{bmatrix} \tag{3.33}$$

where

$$\begin{aligned} C_1 &= \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2) - \bar{x}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \\ C_2 = C_3 &= \sigma_{XY}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i y_i) - \bar{x}\bar{y} \\ C_4 &= \sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i^2) - \bar{y}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \end{aligned} \tag{3.34}$$

C_1 and C_4 are readily recognised as the variances of the x and y values respectively, taken over all of the pixels in the object. C_2 and C_3 are the covariances, which measure how the x and y pixel values in the object are related. If an object of unspecified shape and size is aligned so that its major axis (i.e. its longest medial line in the sense of

minimising pixel moments) lies along the line $y = x$, then it will be found that the term $\frac{1}{N}\sum_{i=1}^{N}(x_i y_i)$ will have a greater value than at any other possible orientation. On the other hand, if the major axis is aligned to a co-ordinate axis such as the x-axis, then y pixel positions will be essentially independent of their corresponding x pixel positions, and the term $\frac{1}{N}\sum_{i=1}^{N}(x_i y_i)$ will be very close in value to $\bar{x}\bar{y}$, yielding covariances very close to zero.

The Hotelling transform aims to rotate the major axis of the object precisely so as to achieve the condition $C_2 = C_3 = 0$, thus ensuring the desired alignment. This is equivalent to diagonalising the covariance matrix, which can be done efficiently by computing its eigenvalue decomposition. The characteristic polynomial of the matrix is

$$\begin{aligned} \det(\mathbf{C} - \lambda\mathbf{I}) &= \begin{vmatrix} C_1 - \lambda & C_2 \\ C_3 & C_4 - \lambda \end{vmatrix} \\ &= (C_1 - \lambda)(C_4 - \lambda) - C_2 C_3 \\ &= \lambda^2 - (C_1 + C_4)\lambda + (C_1 C_4 - C_2 C_3) \end{aligned} \tag{3.35}$$

and so the larger eigenvalue is given by

$$\lambda_1 = \frac{(C_1 + C_4) + \sqrt{(C_1 + C_4)^2 - 4(C_1 C_4 - C_2 C_3)}}{2} \tag{3.36}$$

The homogeneous equation for the system is expressible as an augmented matrix, which reduces as follows:

$$\begin{aligned} [\mathbf{C} - \lambda_1\mathbf{I}\,|\,0] &\sim \left[\begin{array}{cc|c} C_1 - \lambda_1 & C_2 & 0 \\ C_3 & C_4 - \lambda_1 & 0 \end{array}\right] \\ &\sim \left[\begin{array}{cc|c} 1 & \frac{C_2}{C_1 - \lambda_1} & 0 \\ C_3 & C_4 - \lambda_1 & 0 \end{array}\right] \end{aligned} \tag{3.37}$$

from which the eigenvector $\mathbf{e_1}$ with unit x-component value that corresponds to λ_1 (and which must satisfy $\mathbf{C}\mathbf{e_1} = \lambda_1\mathbf{e_1}$) is seen to be

$$\mathbf{e_1} = \begin{bmatrix} 1 \\ \frac{\lambda_1 - C_1}{C_2} \end{bmatrix} \tag{3.38}$$

This eigenvector denotes the current orientation of the major axis with respect to the co-ordinate axes of the image. Thus, the current angle of the major axis of the object, relative to the horizontal x-axis of the image, may be found as

$$\theta = \arctan\left(\frac{\lambda_1 - C_1}{C_2}\right) \tag{3.39}$$

(a) Original leaf orientation (b) After Hotelling transform

Figure 3.23: **An illustration of the rotating capabilities of the Hotelling transform**

so that, knowing θ, a clockwise rotation through this angle will align the object's major axis with the image's x-axis. In the case where it is also desired to restore the centroid of the aligned object to some standard position (x_o, y_o), the entire rotation and translation to the new co-ordinates may then be defined for the output object (x_h, y_h) as

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \left(\begin{bmatrix} x_h \\ y_h \end{bmatrix} - \begin{bmatrix} x_o \\ y_o \end{bmatrix} \right) + \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \tag{3.40}$$

Figure 3.23 shows the results of implementing the hotelling transform on the image of a leaf which was originally photographed at a random orientation.

3.8 Fourier descriptors

Numerous techniques exist for characterising the outer boundary of an object within an image — these include run-length encoding, chain code, fractal dimension and the use of derived shape parameters such as circularity, form factor, extent, compactness, roundness, *et cetera* [59]. The method chosen in this project, because of its general na-

ture and its power in discriminating between the gross shape features and the detailed outline of an object employs the *Fourier descriptors* of the outline.

Each position (x,y) on the object's outer boundary can be represented by a complex number $x+jy$, or may be identified with an angle θ which is the angle which the line joining (x,y) to the object's centroid makes with the image's x-axis. In images of leaves, one occasionally finds positions where a value of θ may correspond to more than one boundary point, especially in a damaged section of the leaf; but, given that a unique point is chosen for each value of θ (as will be the case in this project), the object boundary can then be described by a complex function $f(\theta)$. This function is defined over the range $(-\infty,\infty)$ and, because the leaf outline is a closed contour, it is periodic with period 2π. Since x and y both depend on θ, the function may be written

$$f(\theta) = x(\theta) + y(\theta) \tag{3.41}$$

or, if one envisages traversing the boundary once in every time period $t = T$, so that $t = \frac{T\theta}{2\pi}$, then

$$f(t) = x(t) + y(t) \tag{3.42}$$

The periodic function $f(t)$ has a Fourier transform given by

$$\mathcal{F}\{f(t)\} = F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \tag{3.43}$$

which (because $f(t)$ is periodic) consists of a sequence of discrete weighted impulses at values of ω that are integer multiples of $\frac{2\pi}{T}$. If the Fourier transform is derived from a *sampled* version of $f(t)$, such as

$$f_s(t) = f(t)\delta_T(t) = f(t)\sum_{n=-\infty}^{\infty}\delta(t-n) \tag{3.44}$$

in which one pixel sample per time unit (T samples per period) has been obtained by multiplying $f(t)$ by the Dirac delta train $\delta_T(t) = \sum_{n=-\infty}^{\infty}\delta(t-n)$, then the Fourier transform $F_s(\omega)$ will be both periodic and discrete [49], consisting of T impulses repeated periodically. The weights of these T impulses are known as the *Fourier descriptors* of the boundary $f(t)$. An alternative method for deriving the Fourier descriptors is to take the discrete array of sampled values of $f(t)$, denoted g_r for $r = 0,1,2\ldots T-1$, and apply the Discrete Fourier Transform (DFT):

$$G_u = \frac{1}{T}\sum_{r=0}^{T-1} g_r e^{\frac{-j2\pi ur}{T}} \quad \text{for } u = \frac{-T}{2}+1,\cdots -1,0,1,\ldots\frac{T}{2} \tag{3.45}$$

The values of G in the output array of this transform correspond to the weights of the impulses of $F_s(\omega)$, and are therefore also the complex Fourier descriptors of the boundary function $f(t)$.

The boundary function $f(t)$ is a finite power signal, for which the coefficients of $F(\omega)$ decay to zero like $\frac{1}{\omega}$ or faster. This allows $f(t)$ to be treated as if it were a band-limited signal, which can adequately be recovered from the sampled version $f_s(t)$ provided that the number of samples, T, is large enough to satisfy the Nyquist sampling criterion. All such sampling in this project was done at sufficiently high frequency as to ensure that the effects of aliasing were negligible.

The inverse Fourier transform

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega \tag{3.46}$$

may now be used to recover a *continuous* representation of the boundary, denoted here as $f(t)$ and arbitrarily similar in shape to the original $f(t)$, from the Fourier descriptors contained in $F_s(\omega)$. This is achieved by creating from the periodic frequency domain function $F_s(\omega)$ a non-periodic function $F(\omega)$ by windowing $F_s(\omega)$ symmetrically and retaining only the band-limited portion centred around zero. Then, with

$$F(\omega) = F_s(\omega)\text{Rect}\left(\frac{T\omega}{2\pi}\right) \quad \text{where} \quad \text{Rect}\left(\frac{\omega}{\Omega}\right) \equiv \begin{cases} 1 & |\omega| < \frac{\Omega}{2} \\ 0 & |\omega| > \frac{\Omega}{2} \end{cases} \tag{3.47}$$

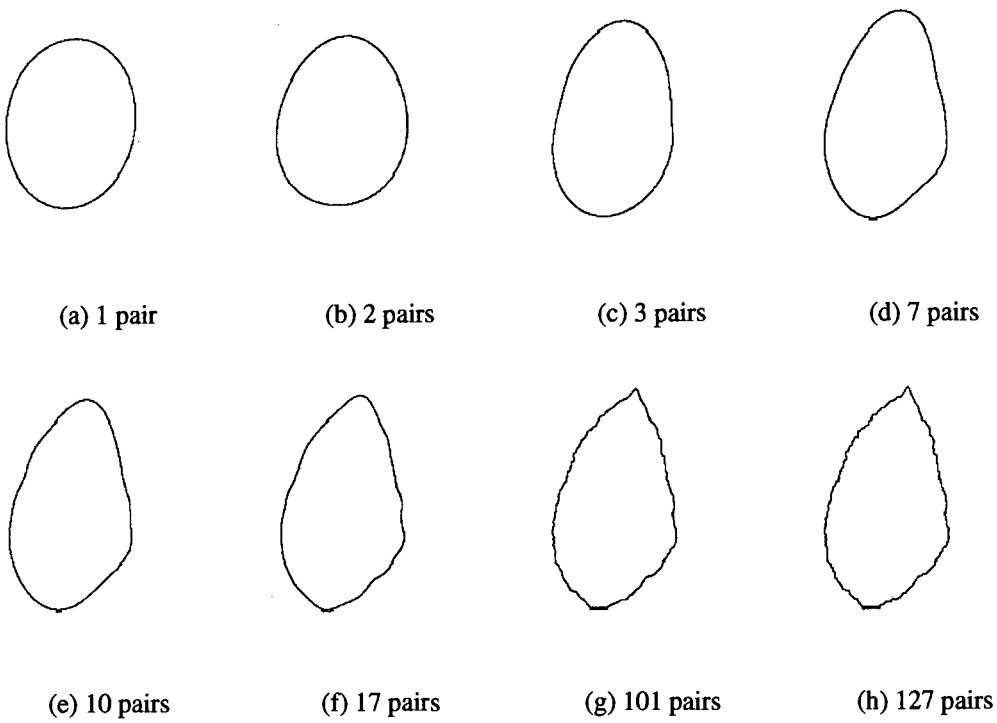
f(t) can be recovered as

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega \approx \sum_{u=-\frac{T}{2}+1}^{\frac{T}{2}} G_u e^{\frac{j2\pi ut}{T}} \tag{3.48}$$

and the inverse DFT may be applied with negligible error, since $f(t)$ was adequately sampled and $F(\omega)$ effectively band-limited.

Figure 3.24 shows the results of this reconstruction for various numbers of Fourier descriptor pairs, in the case of a leaf outline which was sampled with $T = 256$ samples. These reconstructions illustrate how the lower order Fourier descriptors carry information about the gross size and shape of the leaf, while the higher order terms describe the fine detail in the leaf outline.

Besides this, Fourier descriptors have several interesting and useful properties. For example, it may be noted that the Fourier descriptor G_0 is the complex number that



(a) 1 pair (b) 2 pairs (c) 3 pairs (d) 7 pairs

(e) 10 pairs (f) 17 pairs (g) 101 pairs (h) 127 pairs

Figure 3.24: **Reconstructions from various numbers of Fourier descriptor pairs**

represents the *centroid* of the object, and that *translation* of the object within the image can be achieved by varying G_0 and then taking the inverse Fourier transform. The first pair of Fourier descriptors, G_1 and G_{-1}, combined with G_0, carry information about an ellipse whose boundary points are contained in the array

$$g_r = G_1 \left[\cos\left(\frac{2\pi r}{T}\right) + j\sin\left(\frac{2\pi r}{T}\right)\right] + G_{-1}\left[\cos\left(\frac{2\pi r}{T}\right) - j\sin\left(\frac{2\pi r}{T}\right)\right] + G_0 \tag{3.49}$$

Although this ellipse is not the best fitted ellipse (in the least-squares sense) to the leaf outline, nor is its major axis exactly aligned to the major axis of the leaf, nor is its area equal to the area of the leaf, it is nevertheless a first-order representation of the size and shape of the object, representing it to a degree which higher order descriptors will further refine. As figure 3.24 suggests, it does not require many lower order Fourier descriptors to generate a set of features that are quite adequate in the classification of the shape from which they were derived.

The linearity of the Fourier transform allows that multiplication of all of the Fourier descriptors of an object by a constant factor corresponds in the spatial domain to multiplication of the object's length dimensions by the same factor.

Finally, all of the object's Fourier descriptors except G_0 are invariant to translations of the object within the image, and have *magnitudes* which are invariant to rotations of the object. Since the orientation information is stored in the *phases* of the Fourier descriptors, rotation of the object boundary by an angle θ in the spatial domain corresponds to multiplication of all of the Fourier descriptors of the object by the constant complex rotation factor $e^{j\theta}$.

Chapter 4

Leaf Data Acquisition and Preprocessing

4.1 Introduction

The successful segmentation and accurate analysis of objects within digital images may depend critically upon the quality of *acquisition* of the images being processed. Excessively low or high resolution, inadequate illumination or contrast, blur due to poor focusing or camera movement, and overlapping objects within the image can all greatly lengthen the time taken by (and increase the difficulty of) image processing algorithms. Beyond a certain point, such acquisition defects may even make object segmentation impossible. In some image processing applications, such as the identification of a motor vehicle's number plate from a road-side speed camera photograph for instance, there may be little control on the focus, angle or illumination of the object within the image, even though the optical system of the camera is of a reasonable quality. This is unavoidable in a situation where the image must be acquired quickly and the object (the car) is moving with undetermined velocity. However, where one has full control of the objects to be imaged, as in this project, it is desirable to do as much as possible to ensure that information is not being unnecessarily lost or degraded through poor initial acquisition of the images.

In cases where a number of images are to be submitted to a single algorithm, for example to gather comparable data across an entire batch of objects, it is also very desirable to ensure that the objects within the images are presented as consistently as

possible. This means not only that every image should have been acquired (as far as is practicable) under similar conditions of scale, rotation and lighting, but also that such unavoidable differences between images as can be normalised by image processing should be remedied before further processing and data gathering takes place. This is the role of *preprocessing*.

This chapter explains how images of leaves were acquired and preprocessed for use in this project, and it describes the measures that were taken to ensure that high quality and consistent data was available for the segmentation and feature extraction stages which followed.

4.2 Selecting leaves

Every tobacco leaf which was photographed and used as image data in this project had previously been sold at auction at the Zimbabwe Tobacco Sales Floors (TSF), a section of which is shown in figure 4.1(a). This means that not only had it been graded

(a) Tobacco Sales Floors (b) A hand of tobacco

Figure 4.1: **Tobacco bales at auction**

by the farmer who harvested, cured and baled it, but that the bale which contained it had been scrutinised and correctly graded according to official Zimbabwean standards by a professional grader employed by the TSF. Furthermore, since the bale had already been sold when the sample leaves were removed from it, the bale grading had implicitly been accepted as correct by the buyer, who is also an expert in this regard!

Figure 4.1(a) shows several lines of bales which had just been auctioned at the TSF, awaiting the re-sewing of their hessian baling material prior to transport to their purchasers' warehouses. Just before the re-sewing, a hand of tobacco was removed (see figure 4.1(b)) from occasional bales for use as material for data in this project. This was done by the Chief Classifier (Flue Cured Tobacco) of the TSF, who then inspected each hand and tagged it with a label bearing the official grademark of the tobacco. Over the course of about a week, he removed several dozen hands representing a very wide range of tobacco grades, some of which are shown in figure 4.2(a). Within each

(a) Hands of various grades (b) Inspecting individual leaf

Figure 4.2: **Tobacco samples removed from bales**

hand, it was possible to inspect the occasional leaf, as in figure 4.2(b), to ensure that it conformed to the grademark indicated on the corresponding tag. Since these assessments were made by an expert under the excellent lighting at the TSF and again in a dedicated and well-lit tobacco display room, it is believed that the grademark assigned to these hands of tobacco was as accurate as humanly possible.

Altogether, between two and four hands of tobacco for each grademark were identified in this way, and there were 45 different grademarks represented, spanning the most commonly seen flue-cured tobacco grades. The number of leaves per hand varies widely depending on the leaf weight and quality, but is typically between 15 and 30.

4.3 Preparation of leaves

Having been drawn from the bales at the TSF, the hands were then taken to the grading shed of Kutsaga Research Station just outside Harare. The Tobacco Research Board, which operates Kutsaga, has outstanding facilities for the handling and accurate grading of tobacco, and the work for this project was done in the area shown in figures 4.3(a) and (b). With the help of a Technical Assistant of 6 years' experience and a

(a) Grading tables and lights (b) Area used for conditioning

Figure 4.3: **Interior of the grading facilities at Kutsaga**

Grader who had been handling and grading tobacco for 36 years, individual leaves were now selected for photographing. The leaves were first checked individually to ensure that they had all been correctly graded, and occasionally leaves were rejected because they appeared to merit a different grademark from the tag on their hand.

Table 4.1 outlines the numbers and types of leaves which were selected for use in studying leaf colour classification. Another set of leaves, intended for use in studying plant position classification, was selected at the TSF about a year later than this first set. These are tabulated in table 4.2. As these tables show, a total of 870 leaves were selected for imaging purposes. Although the leaves still exhibited visible variability of colour and shape within each colour and plant position class, they were believed, by this stage, to be as correctly individually graded as was reasonably possible.

Cured tobacco leaves are quite pliable, but it was found that few leaves would readily lie completely flat. To obtain a consistently-focused image which gave a faithful impression of the leaf's shape and size, it was essential to ensure that each leaf was spread out entirely flat prior to photographing it. An informal test showed that the ap-

Leaves selected for use in colour classification						
Colour	Grademark	Factors	No. of leaves	TOTALS		
Pale lemon	L2E	-	65	65		
	L3E	-	15	15	80	
Lemon	L2L	-	10			
	L2L	A	15			
	L2L	F	10			
	L2L	FA	15			
	L2L	G	10	60		
	L3L	-	10			
	L3L	A	15			
	L3L	F	10			
	L3L	FA	15			
	L3L	G	10	60		
	L4L	-	10			
	L4L	F	10			
	L4L	G	10	30		
	L5L	-	10			
	L5L	F	10			
	L5L	G	10	30	180	
Orange	L2O	-	10			
	L2O	F	10			
	L2O	FA	15			
	L2O	G	10	45		
	L3O	-	10			
	L3O	F	10			
	L3O	FA	15			
	L3O	G	10	45		
	L4O	-	10			
	L4O	F	10			
	L4O	G	10	30		
	L5O	-	10			
	L5O	F	10			
	L5O	G	10	30	150	
Light Mahogany	L2R	-	20	20		
	L3R	-	20			
	L3R	G	15	35		
	L4R	-	20			
	L4R	G	15	35		
	L5R	-	20			
	L5R	G	15	35	125	
Dark Mahogany	L3S	-	55	55		
	L4S	-	25			
	L4S	G	15	40		
	L5S	-	20			
	L5S	G	10	30	125	660

Table 4.1: **Summary of the leaves selected for colour analysis**

Leaves selected for use in plant position classification				
Plant position	Full grademark	No. of leaves	TOTALS	
Primings	P1LFA	35	35	
Lugs	X1E	35	35	
Cutters	C3LF	35	35	
Leaf	L1E	28		
	L2E	7	35	
Smoking Leaf	H3O	35	35	
Tips	T2RA	35	35	210

Table 4.2: **Summary of the leaves selected for plant position analysis**

parent width of a leaf can easily be reduced by about 15% if it does not lie completely flat — an error which would make shape analysis impossible if it were not avoided.

(a) A typical steambox (b) Conditioning a hand

Figure 4.4: **The use of steam for conditioning leaves before laying them flat**

Figure 4.5: **Ready to film**

The pliability of cured tobacco leaves is improved by *conditioning* them. This simply involves exposing them briefly and evenly to steam, which is a standard procedure in grading sheds, where correct moisture content is essential to the baling of a crop which must leave the farm in a prime state of readiness for potential buyers. For this purpose, grading sheds have a *steambox* such as the one shown in figure 4.4(a). In a steambox, steam is piped to a small water-collecting drum and is then percolated evenly to the air above through hessian and wooden slatting. A hand of tobacco can be conditioned in the steam in a few seconds, as shown in figure 4.4(b). This requires considerable skill and experience, since a leaf which has been over-conditioned may exhibit water staining and a darkening of colour which, for the purposes of this project, would be most undesirable because it might affect the accurate colour classification of the leaf image. A perfectly

conditioned leaf, however, can be made to lie very flat, even if it was formerly rather brittle and shattery due to disease or poor handling in the curing process. Figure 4.5 shows a broken leaf of very poor quality that has been laid flat ready for photography. As with every other leaf photographed in this project, this leaf has been laid onto a clean white backing board and is accompanied by a scale calibrator (in this case some 2cm×2cm squares, although sometimes a small plastic ruler was used), and by a card stating the grade of the leaf (*L5OF* in this case) and the number of the leaf image (No. 8) within the batch of similar *L5OF* leaves.

4.4 Photographing the leaves

The next stage in the process of image acquisition was the photography itself. It was considered very important to ensure that the leaves were correctly and sufficiently illuminated prior to photographing. This meant trying to recreate the lighting intensity and colour under which a human grader would be judging tobacco, so as to make the output of the machine classifier as commensurate as possible with the opinion of the human grader. It also involved avoiding shadows or bright spots such as would be caused by directional lighting or specular reflections.

Fluorescent tubes have been in use in tobacco grading for about 40 years. When they first became available, there was some variation in opinion as to the desirable light intensity on the grading table, but a standard system of having one or more fluorescent colour-matched daylight tubes of 4ft (1.22m) in length situated 4ft (1.22m) above a grading table of rectangular dimensions 4ft×5ft (1.22m×1.52m) seems to have been generally settled upon in the early 1960s [71, 64]. It was found that grading in the light of fluorescent tubes gave more reliable results than the previous system of grading in natural light near or under skylight windows.

Indeed, Gooch, in his doctoral dissertation of 1962 [33], found that both the grading and the price of tobacco was affected by the brightness of the weather in the old naturally-illuminated auction warehouses in Kentucky. United States federal graders were found to have ascribed higher grades and buyers to have bid higher prices on bright sunny days than they had on duller days for exactly the same quality of tobacco. Gooch concludes his work by recommending the standardisation of light sources in grading rooms and warehouses, and the exclusive use of artificial lights.

The current standard in Zimbabwe, as described by Akehurst, calls for well-diffused artificial light of luminous energy around 323 lumen/m^2 over the grading table, preferably supplied by colour-corrected fluorescent tubes only, but with natural roof lights of no more than 10% of the floor area or side lights of no more than $12\frac{1}{2}$% of floor area if absolutely necessary [3]. Direct sunlight is completely unacceptable, and, if at all possible, grading should take place under the artificial light only.

Accordingly, the photographs in this project were all taken under a bank of sixteen 40 watt 4ft (1.22m) Astra^{TM} fluorescent strip lights. The fluorescent tubes were marked as "B.2.1 colour match No. 55", which designates one of the daylight tube types that has been specifically recommended for tobacco grading use [71]. A tripod was set up under these lights as shown in figure 4.6(a), and each leaf to be photographed was then placed under the tripod, as figure 4.6(b) illustrates.

(a) Tripod under lights behind grading table (b) Tripod configuration

Figure 4.6: **The lighting and tripod configuration for photographing the leaves**

In this configuration, the fluorescent tubes were 2.0m above the leaf, and the camera was rigidly attached to the tripod, placed so that the leaf was 900mm below its focal plane. Great care was taken to ensure that neither the photographer nor the tripod cast any shadow over the leaf while it was being photographed. In the case of the leaves photographed to study plant position classification, the distance from the camera's focal plane to the photographed leaf was increased to 1177mm±1mm to facilitate this. Even so, tripod shadows are occasionally to be seen on the outer areas of the card

in some of the images. These were removed in the preprocessing, as discussed later.

The camera used for all of the photography was an Olympus™ OM2N SLR camera body with a standard Olympus 50mm lens. The camera was loaded with Fujichrome SENSIA™ 100ASA colour slide film in rolls of 36 exposures, all from batch number 136103. Fujichrome film was chosen for its excellent colour reproduction and stability over time, while 100ASA slide film was preferred for its fine resolution (lack of graininess) and direct rendering of the image onto positive photo-emulsion, both of which promised very high quality images once the photographs were digitised.

The camera lens was set to F8, and a long time exposure of $\frac{1}{15}$s was employed. This necessitated the use of a shutter release cable to avoid camera movement during the exposure, but had the benefit of allowing an exposure time somewhat greater than the $\frac{1}{50}$s period of the fluorescent tubes, thus reducing any possible effects of their 50Hz flicker. A further advantage of using the shutter release cable was that it enabled the photographer to stand further from the tripod, so avoiding shadowing the leaf.

All 870 leaf photographs were produced in this fashion. The resulting 24 rolls of exposed film (18 for analysis of leaf colour and 6 for plant position) were then professionally developed and mounted as slides to serve as a long-term resource for both this and other tobacco leaf analysis projects.

4.5 Digitisation of the leaf images

The mounted slides were scanned to obtain digital images of the tobacco leaves. This was undertaken with a Nikon™ LS-4500AF film scanner, operated by launching the Nikon Scan TWAIN driver from Adobe Photoshop™. The scanner can accommodate four slides at a time and offers many features in software, including a preview scan, preview image cropping and the choice of focus position on the slide lamina. Each slide was individually focused, and a digitised version of the image in RGB format was then created.

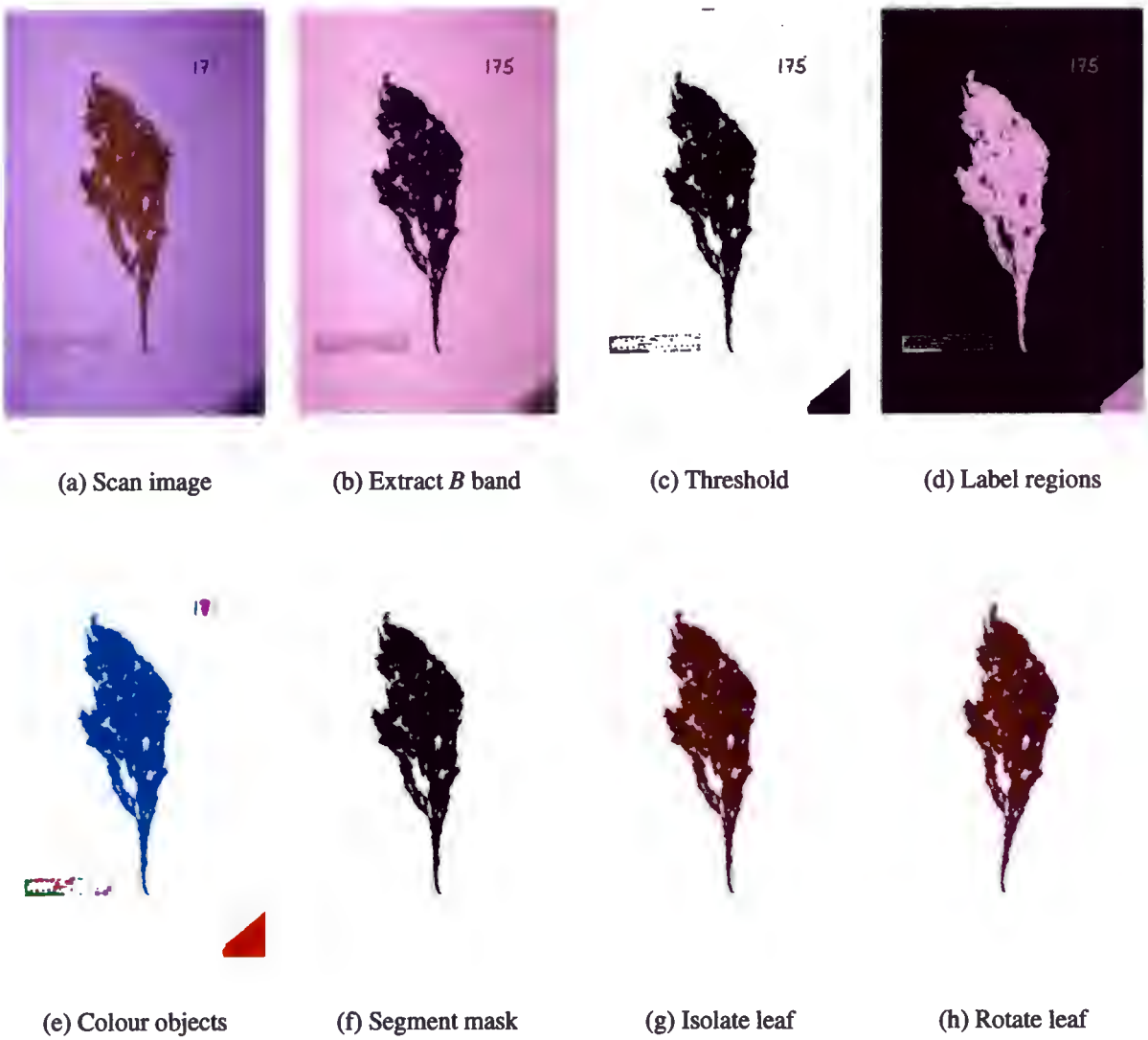
The Nikon scanner is capable of scanning at a variety of output resolutions; and a resolution of 1000 dots per inch (dpi) was chosen for the images to be used in tobacco leaf colour analysis, and 2000dpi for those images destined for the plant position study (where it was felt that highly detailed leaf outline information might be required).

Both of these resolutions are probably higher than was strictly necessary for either part of this project, but it was intended that the digitised images should constitute a permanent resource from which the author and other researchers could draw in the future. Reduction of the leaf images to a resolution that is appropriate to a particular purpose must therefore take place as part of the preprocessing.

The images for tobacco colour leaf analysis were stored as 800×1300 pixel 24-bit colour files in tagged image file format (TIFF). In this case, the TIFF output provided for a lossless compression scheme known as *Lempel-Ziv-Welch* (LZW) compression [59]. This is a dictionary-based technique, by which the image is pre-scanned to identify commonly-occurring patterns of information. A list, or *dictionary*, of these patterns is then constructed, in which a short code is assigned to each of the (much longer) repeated patterns. When the image is stored with the repeated patterns replaced by these short codes, its total size can be much reduced, even despite the fact that it must now be stored together with the dictionary for later decompression. The 800×1300 24-bit colour files, which might have been expected to require $3 \times 800 \times 1300 = 3.12$Mbytes of storage space each, were each accommodated in between 600kbytes and 1.1Mbytes after LZW compression. These compression ratios of 20%-33% are more efficient than the 40%-50% that is typical for the LZW scheme, probably because of the large area of each leaf image that is represented by almost featureless background. The images for use in plant position analysis were scanned at 2000dpi and then stored as 1700×2600 pixel 24-bit colour TIFF image files. These occupied about 4Mbytes of storage space each, again following LZW compression. Finally, all of the scanned images were transferred to writable compact discs, which served as the data repository for the rest of this project.

4.6 Preprocessing of the leaf images

Some of the preprocessing of the scanned images was necessary for only one of the two main algorithms that were developed in this project — that is, either for colour analysis or for plant position analysis only. A description of these specific preprocessing methods has therefore been left until the relevant chapters. This section will give a brief account of the preprocessing that was applied to *every* scanned leaf image, regardless of how it was to be used later.



(a) Scan image (b) Extract B band (c) Threshold (d) Label regions

(e) Colour objects (f) Segment mask (g) Isolate leaf (h) Rotate leaf

Figure 4.7: **Stages in the preprocessing of each scanned leaf image**

The aim of preprocessing in this context was to take each scanned image, such as the image of figure 4.7(a), strip it of its scale calibrator and photograph number, and return a standardised leaf image that would contain only the leaf object from the original image, rotated so that its major axis would be parallel to the vertical axis of the image frame (see figure 4.7(h)). The processing steps in this procedure are illustrated in figure 4.7 and will now be described.

The scanned image (a) was first separated into its component RGB colour bands, and the blue band (B) image was extracted (b). Experience showed that thresholding every blue band image at $B = 170$ would produce a binary representation of the leaf object whose outline was not degraded. Several other objects are visible in the thresholded

image (c), including the residue of a shadow in one corner, the ruler, the leaf number (175), and some minute flecks of dust.

The binary image was now submitted to a *region-labelling* algorithm, which identified the separate objects within the image (there were in fact 69 of them in the case illustrated) and ascribed a different colour to each (image (e)). Region labelling is achieved by traversing the binary image, considering each as-yet-unmarked pixel in turn, and marking both it and every contiguous non-background pixel as belonging to the same object. This is usually done by returning a grayscale image (d) in which a different grayscale pixel value has been given to every pixel of each contiguous object found in this way. The procedure works satisfactorily for labelling up to 255 objects (plus the background "object"), and it can be further extended quite easily. In image (e), for example, which is a 24-bit HSI colour image included here for the purpose of illustration, the different objects have been emphasised by converting their grayscale pixel values to randomly chosen positions around the hue circle and also by giving each pixel a high saturation and intensity.

It was assumed that, by this stage, the leaf would be the largest object in the image. Finding this largest object was a simple matter of finding the grayscale value corresponding to the maximum point of the histogram for the grayscale labelled image. All other objects could then have their grayscale value set to the background value, following which the image was thresholded to a binary format, as shown in image (f). Image (f) can be regarded as a *mask*, which is a faithful representation of the leaf shape (with all its holes) in the original scanned image (a). Since image (a) and image (f) were of the same size in terms of pixel dimensions, multiplication of each pixel value in image (a) by the corresponding value in image (f) (treating black pixels as "ones" and white pixels as "zeros") now yielded image (g), which completed the segmentation of the leaf object from the original image.

For consistency of the stored data, it remained only to rotate the leaf object to be parallel with the vertical image axis. This was done using the Hotelling transform, which showed that (in this case) the leaf's major axis was making an angle of 9.48° with the image vertical. A clockwise rotation of the leaf object by 9.48° about its centroid produced image (h), which is the final form in which the leaf data for this image was stored on CD ROM.

Chapter 5

Some Principles of Machine Vision Classification

5.1 Introduction : the classification problem

Classification is a higher-level function of image processing than any of those that were discussed in chapter 3: a machine vision classifier may incorporate many tools that operate at either the pixel level or on larger-scale structures within the image, but it must also transcend the elementary components of image analysis in its overall operation. The aim of classification in the context of image processing is to assign one of a finite number of *classes* to an image, or to one or more objects within an image. This could be for the longer-term purpose of *object recognition* or of *categorisation*; in each case, the function of the classifier is very similar, with the distinction in the end result deriving from a difference in *interpretation* of the classifier output by its human users. In this project, where each image contains only one object after preprocessing, classification techniques were used to assign each leaf object firstly to one of the five colour categories and secondly to one of the six plant position groups. The specific details of how this was done are left to the next two chapters, while this chapter prepares the way by presenting the relevant theory.

The usual components of a machine vision classifier are shown in figure 5.1, which is an adaptation of the similar diagrams of Low [45] and Jain *et al.* [41]. Several quite different approaches to classification exist, but each can be discussed in terms of this diagram. In every case, classification can be viewed as a succession of *reductions*

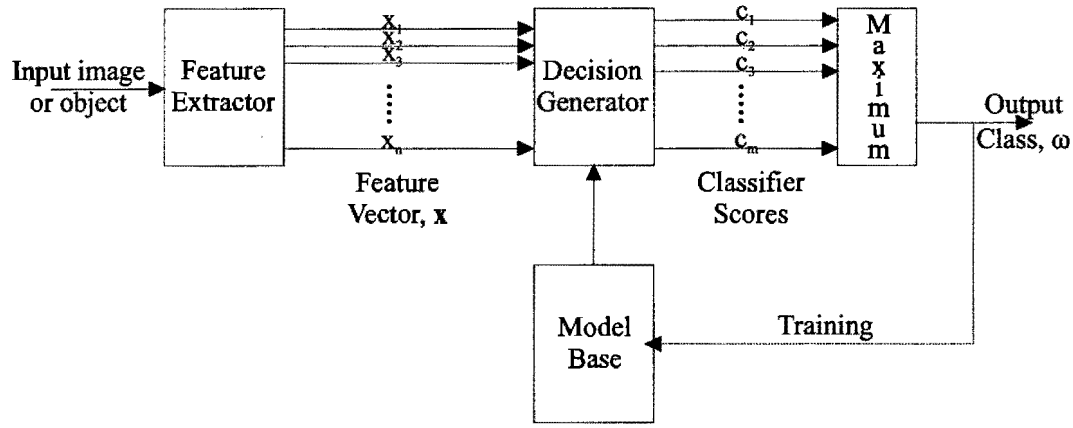


Figure 5.1: **Flow diagram showing the essential stages of classification**

of the information that is stored in an image (or in a segmented part of an image), firstly to a set of (hopefully) diagnostic features, then to a series of scores for all of the possible classes of which the image could be a member, and finally to single class within the usually small set of classes recognised by the classifier.

A detailed treatment of the wider field of statistical pattern recognition cannot be included here, but is the subject of several texts (e.g. [20, 40, 22, 32]) that were of value in deciding upon classification strategies in this project. The rest of this chapter will chart a course through the underlying theory for the tobacco leaf classifiers that were developed in this project. In doing so, it must leave aside such major areas as artificial neural network techniques [9], which offer promise in solving similar classification problems but which were not used in this case because satisfactory results could be obtained by other methods.

5.2 Feature extraction and vector representation.

The concept of classifying an object by feature extraction is very intuitive and is easily presented in a mathematical context that makes its implementation in computer algorithms both natural and straightforward. The image processing techniques covered in chapter 3 offer a variety of ways to extract quantitative measurements from segmented objects within images. In a classifier, these extracted measurements are known as *features*, and there are various types of features that can be derived from image information. For example, some features may be obtained from analysing the

imaged object at the pixel level, and others from measurements of its higher-level geometric structure. Thus, a measure of modal pixel intensity within a region of a leaf gives a *statistical* feature, while the calculation of the leaf's width or the angle of its pointed tip would yield *geometrical* features. One can also distinguish between *global* features whose calculation requires information from every pixel in a region of interest (such as the mean value taken from the red band histogram of an entire leaf image), and *local* features that need only the consideration of a subset of the available pixels (such as leaf boundary length, for example). Some features may be extracted relatively directly from the available image information (for instance, leaf area measured as the total number of pixels in the leaf object), while other *derived* features (such as the Fourier descriptors of the leaf outline) may require considerable intermediate calculation. Finally, a feature which is based upon the relative position or relative orientations of different entities within an image is sometimes known as a *relational* feature [41].

In choosing features that are suitable for a particular classification purpose, there is always some element of human arbitration. As in many of the everyday decisions which we make as human beings, there are usually some criteria in a machine vision classifier which, on their own, provide a fairly good basis for decision; others which are sometimes useful, in conjunction with these, to refine a decision; and still other considerations which either move the decision no closer or, worse, which serve merely to confuse. Whilst there are steps that can be taken to improve and even to optimise the features chosen for use in a machine vision classifier (as discussed later), the initial set of features that are written into the classifier would still have to have been selected by a human programmer, and so will always be subject to the same condition that governs criteria for human decision making — namely that a basis for *perfect* decisions is not always possible to define. This is what gives many machine vision classifiers an *ad hoc* or heuristic flavour [32].

Whatever the case, it is usually possible to take an object in an image (such as a tobacco leaf) and to return a set of values which are the relevant measures for each feature that has been conceived for the object's classification. If this set of values is given an established *order* (e.g. the order in which the features were measured), then it may be treated mathematically as a vector, and is known as the *feature vector* for the object from which it was derived. The feature vector fully characterises the object within the context of the classifier, and will almost always require much less computer

memory for its storage than the original object did. It is the purpose of the feature extractor in a machine vision classifier to condense a (segmented and preprocessed) image or object into a feature vector in this way (see figure 5.1).

It is very often useful to think of a feature vector as a point in space. Thus, if an object is represented by the n-element feature vector $\mathbf{x} = (x_1, x_2, x_3, \ldots x_n)$, then the same object may also be represented by the point which the vector **x** denotes in an n-dimensional space in which the orthogonal co-ordinate axes are assigned to the feature measurands (see figure 5.2). As the next section shows, this makes the method of decision-making easier to treat mathematically, even though the position of objects in n-space becomes rapidly more difficult to visualise as n increases from 2.

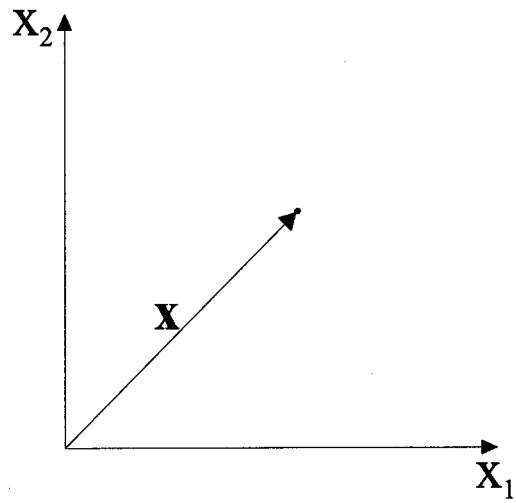


Figure 5.2: **The feature vector x corresponds to a point in (2D) feature space**

5.3 The decision-making process

Once a set of features has been extracted for an imaged object, the next task of the classifier is to make a decision about which class (from the finite set of the classes $\{\omega_1, \omega_2, \omega_3 \ldots \omega_m\}$) the object belongs in. There are numerous ways of doing this, but each involves examining the features and making a decision on the basis of a *model* of the classes, class-boundaries, typical class membership and similarity within classes. The classifier's decision generator applies the model to the feature vector and outputs a set of m scores for each possible class: these scores are finally transferred to another

module of the classifier, which then outputs the class of the input object, most usually on the basis of the maximum (or sometimes the minimum) of these scores.

There are various approaches which designers of classifiers may take to the decision-making process. Perhaps the simplest classifier works on the basis of *direct comparison*. In the case of this project, this would involve passing each leaf image directly through to the decision generator in figure 5.1 (in effect treating each pixel value as an element of the feature vector) and then comparing each pixel to the stored pixels of digitised leaves already in the model. This would imply either a direct subtraction of images, or else a calculation of cross-correlation between images if they were not identically centred and of exactly the same size. The outputs of the decision generator should be made to be low numbers for dissimilar images and a high number for the one image in the model which (almost) exactly matched the input. This kind of classifier would be distinctly clumsy, and is quite unsuitable in this application. It lacks the condensation which is made possible if objects are reduced to features, and it would therefore need massive storage space and possibly a great deal of processing time. In the end, from the point of view of the human user, it would not so much have *classified* the input leaf object as have *recognised* it from a set of pre-existing images in the model base.

Similar objections would hold for *template matching*, in which the input image is reduced to a feature vector in the feature extractor and this is then compared with a set of feature vectors stored in the model, one for each recognisable object. The object in the model whose feature vector is identical to that of the input object is returned as the "class" of the input. Such a classifier would be of value where the inputs were each known to be an exact copy of one of the objects in the model, but it cannot perform on previously unseen input objects, as required in this project.

A third type of decision maker identifies the class of the input object on the basis of a *common property* that is shared by all of the objects in a class and by no other object. A common property might be a discrete fact about the class (e.g. that its objects all have one hole whereas all of the objects in other classes have none), or else it might be a range of values for a measured feature that exclusively indicates the class (e.g. if its length lies between 20cm and 30cm, then it must belong to class number 2). Although in principle this system can classify unseen inputs, it requires an extraordinary degree of separability of the classes within the feature space. Figure 5.3 illustrates in the two-feature three-class case the sort of separability implied, including,

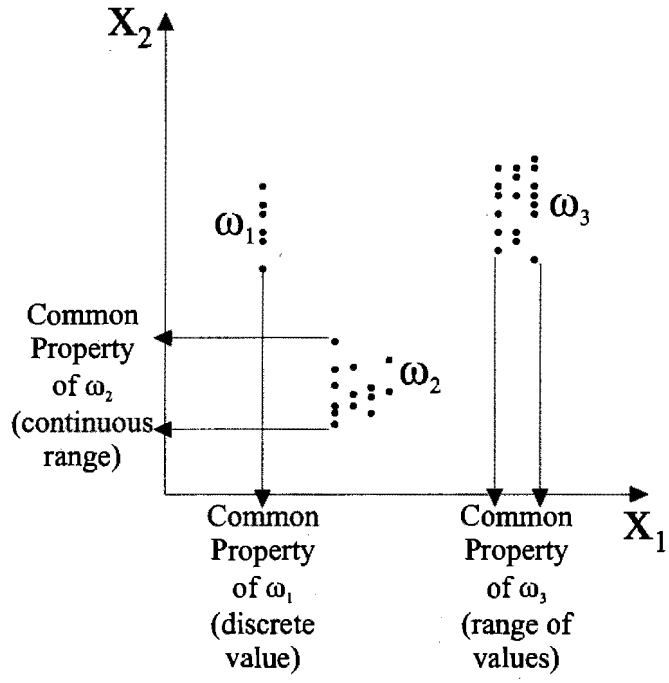


Figure 5.3: **Classification on the basis of common property**

for generality, one feature (x_1) for which the returned values are discrete, and one (x_2) which measures a continuous variable. Common property classification requires that one exclusive property per available class be stored in the model base. In this project it has been found that both the colour and plant position of tobacco leaves exhibit overlap between the class clusters over every feature in the feature space, so a common property approach is not possible.

Instead, problems of this sort need a classifier that holds knowledge of the clustering properties for each class within its model base. This would include knowledge of each cluster's position in feature space, stored for the kth cluster as the vector representing its *centroid*, $\bar{\mathbf{x}}_k = (\bar{x}_{k1}, \bar{x}_{k2}, \bar{x}_{k3}, \ldots \bar{x}_{kn})$, which is shown with the symbol $\odot$ on a cluster diagram as in figure 5.4. It would also hold information about the spread of the cluster, as expressed by its *variance* or, much better in a space where each axis is measured in different units or with varying scales, by the separate variances and covariances of the cluster in each of the co-ordinate axis directions, as specified in its covariance matrix, $\mathbf{C}_k$. Figure 5.4 illustrates the capability of such a classifier. An unknown object, whose feature vector $\mathbf{x}$ is illustrated by the 'X' in the diagram, is to be classified into one of the classes $\{\omega_1, \omega_2, \omega_3\}$. Intuitively, 'X' should belong in the "closest" cluster; but there are several methods for determining which cluster that is.

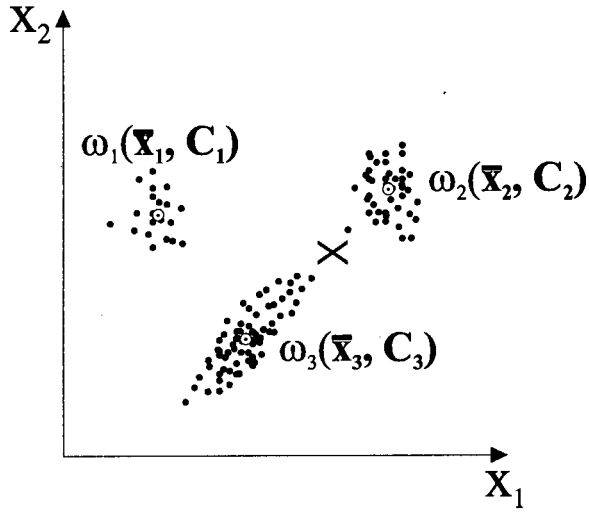


Figure 5.4: **A classification problem that can be solved in several ways**

The *single nearest neighbour* approach [41] assigns 'X' to the same cluster as the nearest sample $\mathbf{u} = (u_1, u_2, \ldots u_n)$ in the n-dimensional feature space. The distance to $\mathbf{u}$ may be measured in the Hamming sense [45] as

$$d = |x_1 - u_1| + |x_2 - u_2| + |x_3 - u_3| + \cdots + |x_n - u_n| \tag{5.1}$$

or by the Euclidean measure

$$d = \left[\sum_{k=1}^{n} (x_k - u_k)^2 \right]^{\frac{1}{2}} \tag{5.2}$$

and in either case the classifier would assign 'X' to class ω_c according to the *minimum distance rule*:

$$d_c = \min_{r=1}^{m} [d_r] \tag{5.3}$$

It is important to notice the assumption here that the units of scale are the same in all dimensions of the feature space. By the single nearest neighbour approach, 'X' would be assigned to the same cluster as the point just above it, which is probably an outlier of class ω_2, but there remains some ambiguity because it is conceivable, looking at the shape of cluster 3, that this point and 'X' both belong in class ω_3.

Evidently, the single nearest neighbour method of calculating cluster membership carries the danger of being distracted by single outliers which are themselves of ambiguous cluster membership, so sometimes a *k-nearest neighbours* rule is preferred [22].

In this case, the nearest k samples to the point 'X' are identified, and 'X' is assigned to the class most frequently represented in those k samples. Despite its nicely democratic flavour, this technique suffers from the arbitrary size of k: as k increases, the accuracy of the classification may tend (up to a point) to improve, but the time complexity of the algorithm will also increase. In terms of figure 5.4, 'X' will be assigned to class ω_2 for $k = 1$, to class ω_3 for $k = 3$, to class ω_2 for $3 < k \leq 42$, and to class ω_3 for $k > 42$, so the method remains ambiguous, and can be distracted by a large more distant cluster even in the presence of a much smaller much closer one. For values of k close to the total number of cluster points, the method becomes meaningless, as it will always assign a new case to the largest class.

The issue of which is the closest cluster to 'X' may be most satisfactorily resolved by considering the position of 'X' in relation to the cluster centroids, and by taking the shapes of the clusters into account as well. The shape of a cluster may be characterised as its degree of spread in the directions of the co-ordinate axes, which is well-represented by its $n \times n$ covariance matrix $\mathbf{C}$, in which the jk^{th} element is

$$\mathbf{C}_{jk} = \frac{1}{Q}\sum_{Q}[(x_j - \overline{x_j})(x_k - \overline{x_k})] \tag{5.4}$$

where Q is the total number of points in the cluster. If $\mathbf{x}$ denotes the multivariate normal distribution of the points in a cluster, then the projection of $\mathbf{x}$ onto a unit-length feature-axis vector, $\mathbf{a}$, has variance $\mathbf{a}^T\mathbf{C}\mathbf{a}$ [22]. Furthermore, the cluster can be drawn as a series of closed contours of equal cluster density, for which the value of the quadratic form $(\mathbf{x} - \overline{\mathbf{x}})^T\mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}})$ is constant. In a feature space with n dimensions, these contours are $(n-1)$-dimensional hyperellipsoid surfaces enclosing the inner cluster. A measure of the distance of a point such as 'X' from the cluster centroid is the local value of the cluster density contour on which 'X' lies, whose square is known as the *squared Mahalanobis distance* and is given by

$$d^2 = (\mathbf{x} - \overline{\mathbf{x}})^T\mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \tag{5.5}$$

This distance measure has the advantage of taking all of the cluster elements and also the shape of the cluster into account, and so it is often the most appropriate of the nearest-neighbour-type classifiers, although it is computationally the most complex. It also has the virtue of being unambiguous. Thus, in figure 5.4, although 'X' appears to be closest to class ω_2 in the single-nearest-neighbour and distance-to-centroid senses, it would be assigned to class ω_3 on the basis of minimum Mahalanobis distance from the class centroids. This makes sense when one looks at the shape of cluster 3.

The action of a classifier can be represented with the use of *decision (or discriminant) functions* [32], so as to formalise what has already been covered in this chapter. In an n-dimensional feature space with m classes, the classification problem is to find m decision functions $d_1(\mathbf{x}), d_2(\mathbf{x}), \ldots d_m(\mathbf{x})$ which have the property that if $\mathbf{x}$ belongs to class ω_i, then

$$d_i(\mathbf{x}) > d_j(\mathbf{x}) \qquad \forall j,\ j \neq i \tag{5.6}$$

The set of outputs from these functions is passed from the decision generator to the final module of the classifier (see figure 5.1), which then assigns the class of the input object according to the decision function with maximum output. The hypersurfaces on which

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x}) = 0 \tag{5.7}$$

are the *decision boundaries* which separate the regions of feature space occupied by each class.

Thus, in a classification problem such as the one posed in figure 5.4, decision functions could be derived using any of the nearest-neighbour-type distance metrics (Hamming, Euclidean, Mahalanobis) discussed above, and decision boundaries could then be generated from these functions. Another way of viewing the problem of finding decision boundaries is to look for a *weight vector* $\mathbf{w} = (w_1, w_2, \ldots w_n, w_{n+1})$ for which scalar multiplication by the *augmented pattern vector* $\mathbf{x}' = (x_1, x_2, \ldots x_n, 1)$ gives zero on the decision boundary, a positive result if $\mathbf{x}$ lies in one class, and a negative result if it belongs to another. In the two-feature case, such a decision boundary is given by

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}' &= \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = 0 \\ &\Rightarrow \quad w_1 x_1 + w_2 x_2 + w_3 = 0 \end{aligned} \tag{5.8}$$

which is evidently the equation of a straight line. This concept generalises easily to flat hyperplanes in higher dimensions, and can be extended, with the inclusion of higher-order powers of features in the augmented pattern vector, to describe curved boundary hypersurfaces as well. In figure 5.5, two decision boundaries, one a straight line and one quadratic, have been generated in this way for the classification problem of figure 5.4. The straight-line classifier has assigned 'X' to class ω_2, whereas the quadratic classifer, which has more weights and so has been able to "learn" to conform

more precisely to the training class boundaries, assigns 'X' to class ω_3. The decision boundaries for class ω_1 are not shown, but could be computed in similar ways.

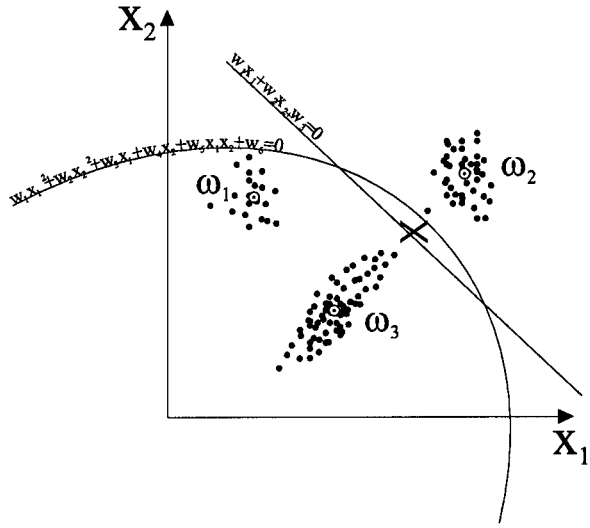


Figure 5.5: **Linear and quadratic weighted decision boundaries**

The detailed mechanics of how decision surfaces are found need not be debated here, but techniques for doing so fall into two categories. In *off-line learning*, fixed prior knowledge of the class boundaries, decision boundary weights or other class characteristics is stored in the model base during the *training* of the classifier, after which all input objects are reduced to feature vectors which are classified according to the fixed "experience" which this model base represents. In *on-line* learning, the classifier has no prior knowledge of the class of any input object, but the class characteristics or decision boundary weight vectors in the model base can be updated automatically on the basis of fresh appraisal of the clustering as more and more input objects are considered. On-line learning has both the advantage of flexibility to change and the disadvantage of liability to be distracted that response to changing input data brings, and it also requires much more elaborate programming than off-line learning. In this project, classifiers were first trained on a data set of objects of known class so as to derive a set of decision functions, and were then put to work to analyse further unseen leaf data, which is referred to as *test data*.

5.4 Bayesian decision theory

Another technique for generating decision boundaries is to plot the decision surfaces as the *loci of equal probability* that a point would lie in either of two adjacent classes. This is a powerful (and indeed optimal [22]) method of separating classes, which in theory works effectively even when there is considerable overlap of the classes in terms of some or all of the measured features. For each of the m available classes, the procedure is to compute the probability that a given input object's feature vector $\mathbf{x}$ could lie in that class. These probabilities are denoted $P(\omega_i|\mathbf{x})$ for classes $i = 1, 2, \ldots m$, and they are termed *a posteriori* conditional probabilities because they rely on prior knowledge both of the candidate feature vector $\mathbf{x}$ and of the classes ω_i. Specifically, if one knows the probability that an object vector, chosen at random from all known objects in the feature space, lies in class ω_i (this is called the *a priori* probability $P(\omega_i)$), and if one knows the probability density function for every feature considering only those objects in class ω_i, $(p(\mathbf{x}|\omega_i))$, then *Bayes' Rule* [9, 43, 44] gives

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{5.9}$$

where the denominator is the total probability density function for a feature vector $\mathbf{x}$, given by

$$p(\mathbf{x}) = \sum_{j=1}^{m} p(\mathbf{x}|\omega_j)P(\omega_j) \tag{5.10}$$

Figure 5.6 lends some insight into how this can be viewed in the two-feature case. Since cluster 1 contains 100 dots and cluster 2 has 50, the *a priori* probability of the next unseen object having a feature vector lying in class ω_1 is considered to be $P(\omega_1) = \frac{100}{100+50} = \frac{2}{3}$, and the corresponding *a priori* probability for class ω_2 is $P(\omega_2) = \frac{50}{100+50} = \frac{1}{3}$. Separate histograms denoting the frequencies of observation for value ranges of the features x_1 and x_2 and for each class, once normalised to unit area, represent the best available estimates of the class-conditional probabilities for the features x_1 and x_2. These probabilities are the component values that may be used to form the vectors $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$. The denominator function $p(\mathbf{x})$ can now be found by combining the class conditional probabilities, using the *a priori* class probabilities as weights. Figure 5.6 shows the components of the denominator function, $p(x_1)$ and $p(x_2)$, calculated in this way.

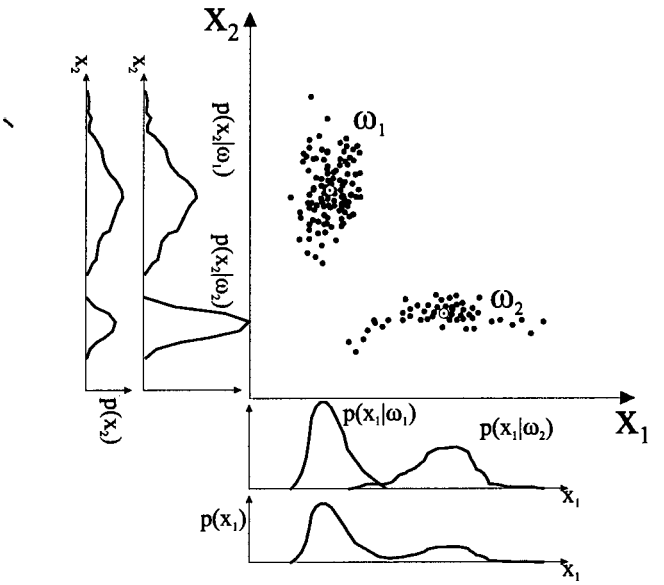


Figure 5.6: **Illustration of the quantities involved in Bayes' Rule**

Once Bayes' Rule has been applied to find a set of *a posteriori* probabilities (one for each possible class) for a candidate feature vector, these probabilities can be passed directly from the decision generator to the final module of the classifier (see figure 5.1), where the input object is assigned to a class on the basis of *maximum a posteriori probability*. It can be shown [22, 32] that the Bayesian probabilistic approach to classification which maximises *a posteriori* probability is an optimal classifier, in the sense of minimising the risk of error. The decision functions for the classifier are simply the *a posteriori* probabilities, so that

$$d_k(\mathbf{x}) = P(\omega_k|\mathbf{x}) \qquad k = 1,2,\ldots m \tag{5.11}$$

and decision boundaries, which may be of quite convoluted shape but which are optimal for the given knowledge, may be generated from these $d_k(\mathbf{x})$ as described in the previous section. It can also be shown [20] that although nearest-neighbour-type classifiers do not explicitly include *a priori* knowledge of the class probabilities, their error rates tend towards the optimal Bayes classifier rate as ever-greater numbers of training set feature vectors are considered. This result supports an intuitive belief that classifying on the basis of minimum Mahalanobis distance from class centroids, which uses *every* piece of available information in classifying an unseen object, implicitly brings *a priori* knowledge of the classes to bear on the problem.

5.5 Classifier design and performance

Gonzalez and Woods [32] make the point that " ... the state of the art in computerized image analysis for the most part is based on heuristic formulations tailored to suit specific problems". This section explores some of the decisions that have to be made in designing a machine vision classifier, and identifies the degree to which general design principles can prevail over *ad hoc* choices for a particular application. It also touches upon methods for measuring and improving the performance of a classifier.

The first issue in classifier design is usually the choice of features to be extracted from the input objects. It is desirable to select features which can be extracted with similar accuracy and fidelity from any input object, so that feature vectors are not misplaced in the feature space due, for example, to a failure properly to measure a feature in an input object. In a machine vision classifier, features should be chosen so as to include all of the major characteristics of the input objects, and should preferably encompass the full range of the objects' attributes, by measuring as many as possible of size, colour, colour variation and shape, wherever these are readily available. The use of derived or relational features is a good idea, if these seem promising; but features should be selected to be as *independent* as possible, so that each feature adds a maximum of new information about an object: somewhat correlated features are inevitable, but high correlations are very undesirable. The author's experience from this and similar projects suggests that any feature that can be seen to differ between classes by a human being, or which seems highly promising in theory, is worth considering in a classifier, provided that it does not impose an intolerable computing or storage burden.

The element of chance in choosing an appropriate feature and, more particularly, in arriving at a set of features which happen to work well together to offer good separability of the classes in feature space, may seem distasteful to those who seek a more "scientific" approach to problem solving. Nevertheless, a careful selection of what seem like suitable features, coupled with a happy synergy as the features operate together in a particular classifier, can yield excellent results — results that may challenge or exceed the performance of a human being in the same task (and one must remember that the workings of the human visual system are themselves far from completely understood). Most machine vision classifiers are created to mimic a human view of reality [40]: in the light of this, it is perhaps not surprising that there is no general theory to govern their design.

Another design issue involves the *number* of features to select for the classifier. A classifier that employs "too few" features will fail to separate the classes effectively, and so may suffer from an unacceptably high error rate. On the other hand, the addition of extra features will not necessarily improve classifier performance, and may well cause it to deteriorate [22, 20, 40]. A high number of features (even independent features) implies a high-dimensional feature space, in which the class clusters may tend to become indistinct and dispersed. A great increase in the number of training pattern points in the feature space may rectify this, but, given a fixed training set, clustering in high-dimensional feature space is quite likely to be poor. In effect, the cluster shapes are distorted by noise that is engendered through the use of features with poor discriminating potential. A high number of features relative to the number of training set cases is likely to lead to a classifier which is "over-trained" on the training set; that is, the classifier becomes too sensitive to the separation of specific cases in the training set as opposed to the class separations of the underlying data. In the extreme case of a classifier with as many features as training set cases, each feature might as well be made identical to an individual training data pattern, which would give the classifier perfect performance on the training data but no ability to classify subsequent unseen test data.

Statistical theory provides, within *discriminant analysis*, a means for rejecting or including a feature from the feature set on the basis of its individual performance in classing the data. In the simple one-dimensional case, the variability among the positions of the class centroids is due in part to the performance of the feature in separating the classes, and due also to the inherent variability within each class. Discriminating criteria which use this fact are of the form

$$L = \frac{S_W}{S_A} \tag{5.12}$$

where S_A is a measure of the variability of the data *among* the classes, and S_W denotes variability *within* the classes [38]. A "good" feature would offer a high spread of the class centroids and only a narrow variance within classes, so L would be a relatively small number. By contrast, a "poor" feature would present much overlap between the classes since the class centroids would be clustered on a scale that compared with the within-class variances, and this would score a high value of L.

In multivariate discriminant analysis, this concept of variability has to be extended to a description of cluster variability in many dimensions. It has already been stated that the variance of multivariate distributions is efficiently expressed by the covariance ma-

trix. Since the diagonal elements of the covariance matrix contain the variances of the features, while the off-diagonal elements constitute the covariances, a transformation of the feature set to a set of derived *factors* can be achieved by *diagonalising* the covariance matrix. The resulting matrix is a covariance matrix for which the new factors are *independent* descriptors of the objects to be classified (because it has zero values for all covariances). This is the *principle axes transform* [40], and it will be noted that it is identical in form to the diagonalisation undertaken in the Hotelling transform (see chapter 3). The remaining diagonal elements of the new covariance matrix give the variances of the derived factors, and any factor whose variance is above an arbitrary threshold may be regarded as a "good" discriminator, while factors whose variances fall short of the threshold may be discarded altogether with very little adverse impact upon the classifier. It is important to note that the principle axes transform does not improve the separability of the classes in the feature space, but that it may, by indicating the way in which the number of features can be reduced, allow for the choice of an optimal subset of the available features which will minimise the computational costs of the classifier.

Multivariate discriminant analysis makes use of matrices which are summations of variability information in covariance matrix form, both within classes and over the whole distribution of the data [44]. The *within-groups sums of squares and cross-products* matrix is denoted as **W**, and its jk^{th} element is calculated as

$$W_{jk} = \sum_{g=1}^{m} \sum_{i=1}^{N} \left[(x_{jgi} - \bar{x}_{jg*})(x_{kgi} - \bar{x}_{kg*}) \right] \qquad j,k = 1 \ldots n \tag{5.13}$$

where g represents the class number, j and k are two particular elements within a class, i denotes one of the total of N elements over all classes, and the asterisks indicate a mean to be taken within a class. The *total sums of squares and cross-products* matrix is a similar measure of variability taken over all of the data, and is denoted **T** with jk^{th} element

$$T_{jk} = \sum_{g=1}^{m} \sum_{i=1}^{N} \left[(x_{jgi} - \bar{x}_{j**})(x_{kgi} - \bar{x}_{k**}) \right] \qquad j,k = 1 \ldots n \tag{5.14}$$

where the asterisks now indicate averaging as j and k range over the entire set of data. A statistic for the overall discriminating power of a feature set, known as *Wilks' lambda* [44, 72], is then given as the ratio of the determinant of the within-groups variability matrix to the determinant of the total variability matrix:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{5.15}$$

The Statistica™ package, which was used extensively in this project to perform discriminant analysis, uses a multivariate analysis of variance (MANOVA) approach to compute Λ. It then calculates the additional separability conferred by adding a new feature to the set, as the multiplicative increment of Λ:

$$\lambda = \frac{\Lambda(\text{after new feature added})}{\Lambda(\text{before new feature added})} \tag{5.16}$$

This ratio is known as the *partial lambda* for the new feature: λ always lies between 0 and 1; and if near 0 it indicates a feature whose use adds very little discriminatory power to the classifier's feature set, while if near 1 it denotes a very valuable feature for classification purposes. Since the statistical significance of a given partial lambda depends on the number of degrees of freedom that the total number of cases (N), the number of classes (m) and the number of already existing features (n) give to the problem, the meaning of λ is interpreted through the use of the *F-statistic* [72, 44]

$$F = \left(\frac{N-n-m}{n-1}\right)\left(\frac{1-\lambda}{\lambda}\right) \tag{5.17}$$

Statistica allows a threshold value of F to be selected so that features will be rejected from the set if their discriminatory performance is found in this way to be significantly poorer than that of other features already in the feature set.

Closely related to the number of features is the issue of the number of input objects that should be chosen for the training data set of a classifier. The larger this number is, the more accurate will be the statistical picture of the clustered distribution of the data in feature space which is held in the model base. Nevertheless, the amount of training data cannot increase without bound: it will inevitably be limited by the quantity of relevant data that can be gathered and labelled; it will have to conform to the capacity of the model base, and it must also not be so huge as to challenge the computing power available for the classifier's program.

The performance of the classifier is eventually measured with the use of a *test set*, which is a set of input object data whose true classes are known but are hidden from the classifier. The classifier's performance may be reported simply as the percentage of the test set cases that are correctly classified after the classifier has been trained and then tested. This raises the issue of the size of the test set. The larger the test set, the lower will be the error on the assessment of classifier performance, so a substantial test set is clearly more desirable than a small test in which a random classification error might much affect the quoted classifier performance. However, this should not

be at the expense of the training set size, especially if there is only a limited number of pre-labelled data objects available.

If the available data is limited so that the training set does not have the same class proportions as the true proportions of the classes in the "real world", then the performance results of the classifier must be interpreted with great care if the same classifier is subsequently to be used to assign classes to unseen "real world" data. Classifier testing in this project was done by deriving the *a posteriori* probabilities for each test object across all classes: the class with the maximum *a posteriori* probability was then compared to the actual class of the test object, and the final success rate over all the test objects was quoted as the classifier performance.

If, however, the training data class proportions are different to those of the "real world", then this should be taken into account by adjusting the test performance result to give a quotable performance for later "real world" use. Given an observation $\mathbf{X}$, Statistica calculates the *a posteriori* probabilities for the m classes as the set of numbers $P(\omega_i|\mathbf{X})_T$ for $i = 1,2,\ldots m$, where the T subscript indicates that these probabilities are based on knowledge of the training set data distribution among the classes, but not on knowledge of the "real world" proportions. For a given class, ω_i, the *a priori* probability of a random observation being from that class is written as $P(\omega_i)_T$ for the training set, and as $P(\omega_i)_R$ in the "real world". It is assumed that the training data is at least a fair representation of the "real world" in that the class-conditional probability density function of the observation X is the same in either. This is in accordance with the fact that exactly the same classifier with the same feature set is to be used in the "real world" as was used in the test after training, and it allows one to write:

$$p(\mathbf{X}|\omega_i)_T = p(\mathbf{X}|\omega_i)_R = p(\mathbf{X}|\omega_i) \tag{5.18}$$

Bayes' Rule may now be stated for the two different situations as

$$P(\omega_i|\mathbf{X})_T = \frac{p(\mathbf{X}|\omega_i)P(\omega_i)_T}{\sum_{i=1}^{m} p(\mathbf{X}|\omega_i)P(\omega_i)_T} \quad \text{and} \quad P(\omega_i|\mathbf{X})_R = \frac{p(\mathbf{X}|\omega_i)P(\omega_i)_R}{\sum_{i=1}^{m} p(\mathbf{X}|\omega_i)P(\omega_i)_R} \tag{5.19}$$

which swiftly gives

$$\begin{aligned} P(\omega_i|\mathbf{X})_R &= \left(\frac{P(\omega_i)_R\, p(\omega_i|\mathbf{X})_T}{P(\omega_i)_T}\right)\left(\frac{\sum_{i=1}^{m} p(\mathbf{X}|\omega_i)P(\omega_i)_T}{\sum_{i=1}^{m} p(\mathbf{X}|\omega_i)P(\omega_i)_R}\right) \\ &= G(i)K \end{aligned} \tag{5.20}$$

where $G(i)$ is the left-hand part of the above which is a function of i, and where the *constant* ratio of the two summations is written as K for convenience. Since the classes in feature space are mutually exclusive and exhaustive of the data,

$$\sum_{i'=1}^{m} P(\omega_{i'}|\mathbf{X})_R = 1 \tag{5.21}$$

and so

$$K \sum_{i'=1}^{m} G(i') = 1 \Rightarrow K = \frac{1}{\sum_{i'=1}^{m} G(i')} \tag{5.22}$$

(Here the index variable is written as i', simply to avoid confusion with the main index variable of interest, which is i). From the above, the "real world" *a posteriori* probabilities can now be expressed as

$$P(\omega_i|\mathbf{X})_R = \frac{\frac{P(\omega_i)_R \, p(\omega_i|\mathbf{X})_T}{P(\omega_i)_T}}{\left(\sum_{i'=1}^{m} \frac{P(\omega_{i'})_R \, p(\omega_{i'}|\mathbf{X})_T}{P(\omega_{i'})_T}\right)} \tag{5.23}$$

which are simply the original *a posteriori* probabilities divided by the training data proportions, then multiplied by the "real world" proportions, and finally normalised so that they total 100%.

The class of the new maximum of these adjusted probabilities for each test object is then compared with the object's known class, and a classifier performance score is deduced for each class. The weighted average of these scores (weighted according to the class proportions in the "real world") gives the new performance indicator, correctly adjusted for "real world" usage of the classifier.

This concludes the discussion of the theory of classifier design: the next two chapters will relate how this theory was applied in the classification of tobacco leaf data by colour and plant position.

Chapter 6

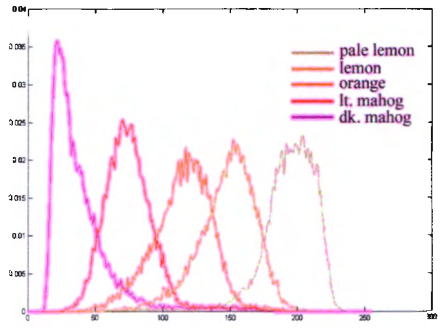
Tobacco Leaf Colour Classification

6.1 Features for colour discrimination

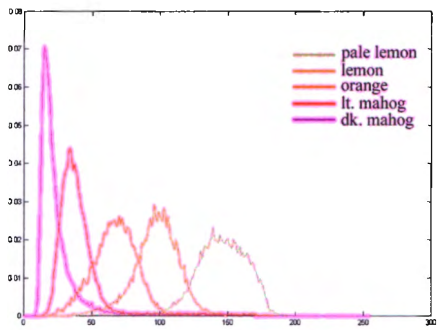
The design of the tobacco leaf colour classifier began with the selection of a set of features that showed some promise in discriminating the official colour grades from tobacco leaf images across the full range of possible tobacco colours. Since the colour of cured tobacco varies from a light greenish yellow to a dark reddish brown, (see figures 2.5 and 2.6), it was conjectured that measurements of the red (R) and green (G) content of a leaf might be good indicators of its colour within the range. There is also a dramatic variation in the darkness of the leaves of different colour grades, and so it was also decided to measure the intensity content (I) of each leaf as another potential colour discriminator. The blue content of the leaves was not measured, since it would be implicitly contained in the R, G and I measurements, via equation 3.5.

Since tobacco leaves exhibit considerable variation in colour and darkness across their surfaces (especially if they are leaves of low quality), it was clear that R, G and I measurements of many pixels within each leaf (preferably including information from *every* pixel within the leaf object) would need to be taken to capture this variation. The mean values, $\overline{R}$, $\overline{G}$ and $\overline{I}$, would then give single-valued representations of the overall colour of the leaf, while the variances, σ_R^2, σ_G^2 and σ_I^2, would be measures of the spottedness, blemish or other regional darkening of the leaf.

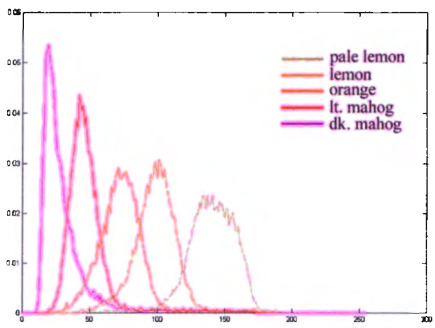
To extract these features, it was decided to derive the R, G and I histograms for each leaf. Because the size of the leaf was considered irrelevant to a determination of its



(a) Red



(b) Green



(c) Intensity

Figure 6.1: *R*, *G* **and** *I* **normalised histograms for all five colour classes**

colour class, and for statistical convenience, each histogram was immediately *normalised* to unit area before features were extracted from it. Normalised histograms for one leaf from each of the five colour categories are shown in figure 6.1, where the variation in the mean values across the classes is easy to see. Since dark blemishes and leaf veins are much more visible in contrast to a light-coloured background than to a

dark-coloured one, the variances of the paler leaf histograms are greater than those of the mahogany leaves, and this is also clearly visible in the graphs of figure 6.1. It was concluded from this that the three means and the three variances all appeared to be promising diagnostics of the colour class.

It was also believed that the higher-order moments of the normalised histograms might carry some discriminatory power. *Moments about the mean* are given by

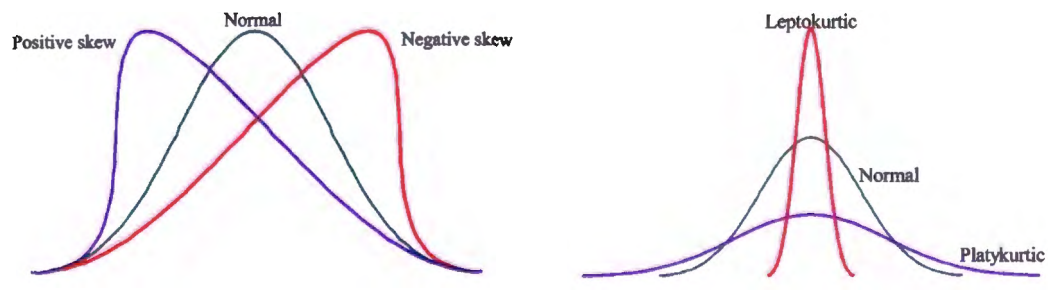
$$\mu_r = \sum (x - \mu_1')^r \phi(x)$$

where $\phi(x)$ is the *probability density function* of the variable x, a function which is well-approximated by the normalised histogram of x in the sense described above. The primed quantity is known as a *moment about the origin*, and may be found from

$$\mu_r' = \sum_x x^r \phi(x)$$

It will immediately be observed that μ_1' is the *mean* of the distribution. Aitken [2] then gives the next three higher moments about the mean as:

$$\begin{aligned}
\text{VARIANCE} : \mu_2 &= \sum (x - \mu_1')^2 \phi(x) \\
&= \mu_2' - (\mu_1')^2 \\
\text{SKEWNESS} : \mu_3 &= \sum (x - \mu_1')^3 \phi(x) \\
&= \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\
\text{KURTOSIS} : \mu_4 &= \sum (x - \mu_1')^4 \phi(x) \\
&= \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4
\end{aligned}$$



(a) Skewness (b) Kurtosis

Figure 6.2: **Illustration of skewness and kurtosis in a distribution**

Figure 6.2(a) shows three probability density distributions, one of which is a *normal distribution*, while the others exhibit *positive skew* and *negative skew*. The potential of *skewness* features such as μ_{3R}, μ_{3G} and μ_{3I} (derived respectively from the R, G and I normalised histograms) to diagnose tobacco colour classes may also be seen from the graphs in figure 6.1, in which the tendency for the probability density functions of the darker leaves to be positively skewed, and for those of lighter leaves to be negatively skewed, is apparent in the R, G and I distributions.

Figure 6.2(b) illustrates what is meant by *kurtosis*, by comparing a normal distribution to one which is *platykurtic* (having a low value of μ_4) and one which is *leptokurtic* (with high μ_4). It was considered possible that the kurtosis values of R, G and I could have potential for colour classification, because the graphs of figure 6.1 tend to range from leptokurtic for dark mahogany to platykurtic for pale lemon (although this is more marked in the cases of μ_{4G} and μ_{4I} than it is for μ_{4R}.

Because the skewness of a distribution tends to perturb the position of its mean (so that the mean is a less good indicator of the unblemished average lamina colour, for example), it was felt that another measure of the central tendency of distributions, which would not be affected in this way, might also offer a useful colour feature. The *modal* values, R_{md}, G_{md} and I_{md}, were therefore also adopted as candidate features, although they were not taken to replace the mean values altogether because, as figure 6.1 shows, modal values are very susceptible to displacement by the random "noise" in the R, G and I distributions derived from tobacco leaves.

Analysis of the colour and intensity distributions of the five leaf colour classes in figure 6.1 therefore produced 15 candidate features for the colour classifier. These are listed in table 6.1. It was recognised that these features are by no means independent, and

$\overline{R}$	$\overline{G}$	$\overline{I}$
R_{md}	G_{md}	I_{md}
σ_R^2	σ_G^2	σ_I^2
μ_{3R}	μ_{3G}	μ_{3I}
μ_{4R}	μ_{4G}	μ_{4I}

Table 6.1: **Fifteen colour classifier candidate features**

figure 6.3 gives an illustration of the cross-correlations of all 15, plus a sixteenth feature, to be discussed in the next section, which was added later. Accessing the small graphs in the figure makes it clear that mean and modal measurements are strongly

positively correlated for each of R, G and I, as could be expected from figure 6.1. $\overline{R}$ is also positively correlated with $\overline{G}$ and $\overline{I}$, and the modes show similar behaviour. Furthermore, the means and modes are quite strongly negatively correlated with the skewnesses in figure 6.3, and there is also a visible relationship in figure 6.1 between the variances and the kurtoses. In fact, there is a statistical tendency for odd-order moments about the mean to be correlated amongst each other and for even-order moments to behave likewise, with successively higher moments adding less and less extra information about the probability distribution from which they were derived. This was what had limited the search for features to the fourth (kurtosis) moments in the first place.

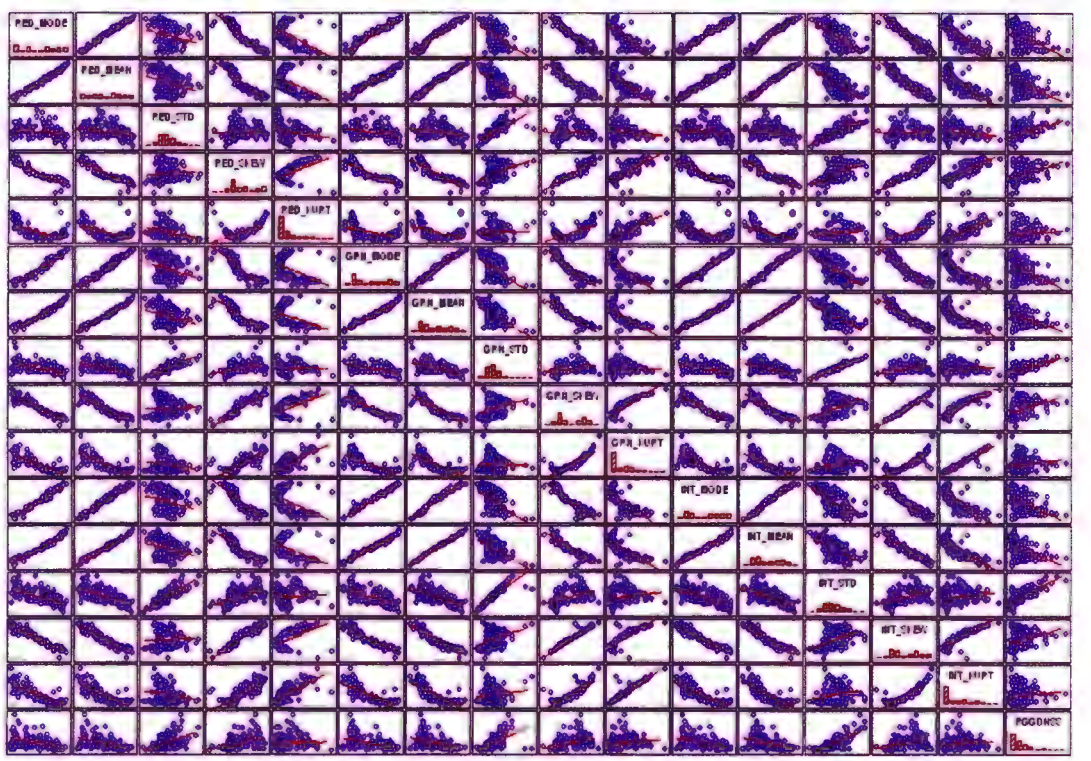


Figure 6.3: **Illustration of the correlations among all sixteen candidate features**

Ordinarily, all of these correlations might have precipitated the immediate removal of many of the candidate features, but in this case all of the features were kept for the time being, firstly because they individually seemed to have desirable properties, secondly because the data for figure 6.1 were based on only five high-quality leaves, and thirdly because it was unknown which of the 15 candidate features would work together as the most effective discriminatory subset following discriminant analysis.

The reduction of the feature set was therefore left until the discriminatory power of each feature had been fully analysed statistically.

It was noted that the candidate features were not measured in commensurate units: whereas the means and modes were measured in the units of the data (pixel brightness in this case), the variances, skewnesses and kurtoses would be quoted respectively in the square, cube and fourth powers of these units. Following Aitken [2], it was decided to remedy this by using the standard deviation as the main measure of spread, and by dividing powers of the standard deviation into the higher-order quantities, to obtain

$$R_{\text{std}} = \sqrt{\mu_{2R}} \quad G_{\text{std}} = \sqrt{\mu_{2G}} \quad I_{\text{std}} = \sqrt{\mu_{2I}}$$

$$R_{\text{skew}} = \frac{\mu_{3R}}{(\sigma_R)^3} \quad G_{\text{skew}} = \frac{\mu_{3G}}{(\sigma_G)^3} \quad I_{\text{skew}} = \frac{\mu_{3I}}{(\sigma_I)^3}$$

$$R_{\text{kurt}} = \frac{\mu_{4R}}{(\sigma_R)^4} \quad G_{\text{kurt}} = \frac{\mu_{4G}}{(\sigma_G)^4} \quad I_{\text{kurt}} = \frac{\mu_{4I}}{(\sigma_I)^4}$$

This resulted in the set of features listed in table 6.2, which were then used in developing the colour classifier.

$\overline{R}$	$\overline{G}$	$\overline{I}$
R_{md}	G_{md}	I_{md}
R_{std}	G_{std}	I_{std}
R_{skew}	G_{skew}	I_{skew}
R_{kurt}	G_{kurt}	I_{kurt}

Table 6.2: **Fifteen features used in the colour classifier design**

6.2 Quality as an extra colour feature

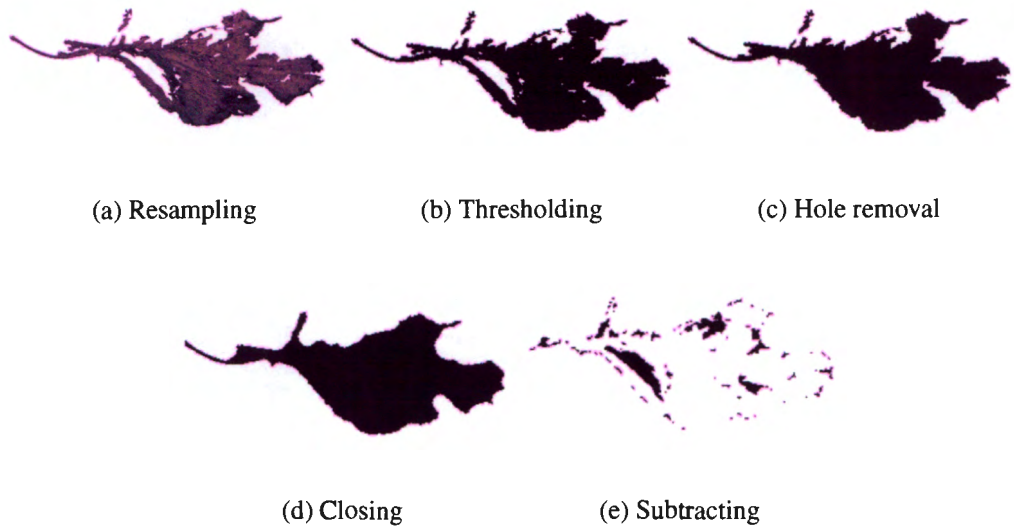
A complication in tobacco leaf colour classification is that, as illustrated in figure 2.7, leaf blemish and the damage associated with low quality may often manifest itself as a darkening of leaf colour. A very poor quality leaf which is, say, lemon in colour may appear, as in figure 2.7(c), very close in overall coloration to a typical orange leaf. Likewise, damaged orange leaves are easily mistaken for light mahogany leaves, and so on. In general, there is a tendency for low-quality leaves to have many darkened areas, and this carries the danger of their being classified into a darker colour category than the underlying colour of the undamaged parts of their laminæ would warrant. In

terms of the histograms of these damaged leaves, the mean is at a lower value and the variance is greater than would be the case for an undamaged leaf of the same colour category. The skewness and kurtosis are also somewhat affected by leaf damage. Figure 6.4 shows a lemon leaf of quality 5 and an orange leaf of quality 2: just from the colour, few but the most experienced graders would be able to tell which was which (in fact, the lemon leaf is the more damaged one on the left).

Figure 6.4: **Lemon leaf of quality 5 easily confused with orange leaf of quality 2**

This problem was addressed by introducing *quality* itself as an extra colour feature, in the hope that adding an extra dimension to the feature space might separate clusters that had overlapped purely because of some low quality leaves which they contained. Leaf quality can, to some extent, be inferred by the degree of tearing or raggedness of the leaf lamina, because it is lamina damage which causes much of the darkening of a poor quality leaf. What was therefore required was a means of estimating the amount of tearing which each leaf had undergone. Figure 6.5 illustrates the sequence of image processing steps that was applied in order to make this estimate.

The leaf was firstly (a) reduced to a horizontal size of 300 pixels without changing the aspect ratio, in order to allow all subsequent steps to run faster whilst not degrading the leaf image enough to reduce the accuracy of the quality estimate unduly. The reduced



(a) Resampling (b) Thresholding (c) Hole removal

(d) Closing (e) Subtracting

Figure 6.5: **The processing steps in extracting the *raggedness* feature.**

image was then (b) thresholded in the blue band at $B = 175$. Next, the background of the thresholded image was region-labelled and then isolated, which resulted in a binary image with the outline appearance of the original leaf, but with all of the holes removed (c). A morphological closing operation was then applied to this binary image, so as to smooth the ragged leaf outline, as shown in figure (d). The structuring element used for this purpose was a circular disc of diameter 18 pixels — small enough for the closing operation to fill the narrow tears in the leaf outline, but not so big as to change the basic shape of the leaf object.

In the final step, the difference image between the closed binary image of figure (d) and the binary silhouette of figure (b) was calculated, so as to arrive at figure (e), which is a representation of the holes and the edge tears in the original leaf object. When the total area of figure (e) is expressed as a percentage of the area of figure (d), one has a measurement which is an estimate of the degree of tearing of the leaf lamina, and hence an estimate of the leaf quality in the context of colour classification. It should be stressed that this percentage is *not* a measure of official grading quality (which includes many subtle leaf surface attributes), but that it does give a suggestion of the degree of darkening that might be expected in a given damaged leaf, on the assumption that tearing is indicative of lower lamina darkening. For the purposes of this project, this percentage was known as *raggedness*, and it was included as a sixteenth feature in the feature vector for every leaf analysed by the colour classifier.

These leaves are listed in table A.1 of appendix A, together with all 16 of the measured feature values for each leaf (rounded to suitable lengths for the purpose of presentation in the table). At the bottom of the table, the feature means are given for each of the five colour classes, and it may be noted that several of them show the systematic tendency to increase or decrease across the class range that has already been alluded to. In particular, the new raggedness feature was found to have an average value of about 1% in pale lemon leaves, rising to nearly 6% in dark mahogany leaves, which gave further promise that it might prove to be a good feature for colour discrimination in its own right.

6.3 The training of the colour classifier

In order to limit the data storage requirements and the processing time for the colour classifier during its development, it was decided to work with a subset of the available leaf images. Thus, 252 leaf images were selected from the data bank so as to include leaves of all colours and qualities; and some leaves that had been graded with the extra factors "ripe" (F) and "green" (G) were also included in order to make this data subset as representative as possible of the tobacco colours that might need to be graded on a farm or at auction. It was decided to use 152 of these 252 images as training data and the other 100 as test data, with this division being made in order to have a large training set and a fairly large test set with a round number of members. The distribution of the 252 images between the colour classes and between training and test data is given in table 6.3. The second column of table A.1 in appendix A gives the colour classes into

Colour	ple	lem	ora	lma	dma	TOTAL
Training set	18	37	31	36	30	152
Test set	12	24	22	23	19	100
TOTALS	30	61	53	59	49	**252**

Table 6.3: **Distribution of the classifier data between training and test sets**

which each leaf had been graded by expert human graders, and the third column shows how the full list of 252 leaves was divided into those selected as training data (column entry 1) and those that would be used later as test data (column entry 2). The remaining columns give all the data, which were extracted for each leaf using MATLAB code

that implemented all of the necessary algorithms in the manner discussed in previous chapters.

The discriminant function analysis package of Statistica was then applied in standard mode to the use of the sixteen features in classifying upon colour. The first outputs of the analysis were the matrices of tables A.2 and A.3, which give the pooled within-groups correlations and total-data correlations for the 16 features of the 152 training data leaves. It is the latter of these two tables which is a condensation of the information shown in figure 6.3. The ratio of the determinants of these correlation matrices was then calculated, and this yielded a Wilks' lambda of $\Lambda = 0.0026475$ for the 16-feature set of the initial model. This can be interpreted as meaning that the model had very high discriminatory potential using these 16 features (the associated F-value was 28.8). Similar calculations were then done by Statistica to derive the entries in table A.4 (see appendix A), which is a summary of the discriminatory potential of each of the features taken separately.

The first column of this table gives the Wilks' lambda for the overall model that would result after removal of the respective feature variable, and the second column uses this information to state the feature's partial lambda. The F-value associated with each partial lambda is given in column 3, and the final column holds the confidence level (in the range $0 \leq p \leq 1$) attached to this F-to-remove value. As the notes for Statistica [72] mention, the numbers in this final column must be treated with caution: they are only really meaningful as a confidence level for the removal of one feature at a time, namely the feature which is currently *least* valuable as a discriminator. Thus, from table A.4, if one wished to reduce the size of the feature set, the first feature to remove ought to be the intensity mode, I_{md}. The tabulated results show that the removal of I_{md} would give the least increase in Λ (from 0.0026475 to 0.00267 which corresponds to a partial lambda of $\lambda = 0.993$). Values of λ may vary from 0 to 1: a value near 0 indicates a feature which is vital for the effective operation of the feature set; a value near 1 denotes a feature which is serving no discriminatory purpose. The I_{md} feature had by far the lowest associated F-value, and the confidence level that can be attached to the hypothesis that I_{md} should be the first feature to be removed was 92.6%.

In deciding how to adjust the feature set, there exists the dilemma that a feature with low correlation with other features may have this low correlation either because it is bringing genuinely independent, fresh information to the classifier, or because it is simply uncorrelated "noise". A second difficulty with deciding which features to

include or remove is exemplified by the case of R_{md}. The second feature suggested by discriminant analysis for removal was R_{md} (presumably because of its high correlation with $\overline{R}$ and several other features), and yet in another analysis in which Statistica was used to compile a subset of useful features by *addition*, starting from no features at all, R_{md} was the first one of the 16 features to be selected for inclusion! The orders in which two separate stepwise analyses indicated that features ought to be removed from the full set or added starting from none are give in table 6.4. It was found by extended experiments with Statistica that feature sets selected by forward stepwise discriminant analysis differ in constitution from those left after backwards stepwise reduction, but do not differ significantly in their performance on test data in this project.

No of features	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
By removal of	I_{md}	R_{md}	$\overline{I}$	Rggdnss	I_{std}	R_{std}	G_{std}	G_{md}	R_{skew}	I_{kurt}	I_{skew}	G_{kurt}	R_{kurt}	G_{skew}	$\overline{G}$	$\overline{R}$
By addition of	R_{md}	G_{md}	R_{kurt}	G_{kurt}	$\overline{G}$	$\overline{R}$	I_{kurt}	G_{std}	R_{skew}	G_{skew}	I_{skew}	R_{std}	I_{std}	Rggdnss	$\overline{I}$	I_{md}
No of features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Table 6.4: **Order of removal and addition of features suggested by stepwise analyses**

The question of how to optimise a feature set by reducing the number of features in it can be debated almost endlessly, and in this project it was found that there were many feature sets of different sizes which, after training of the classifier, scored reasonably well on the test set. It was decided to reduce the feature set to the number of features below which there was evidence that the classifier performance might be compromised. The advantages of having fewer features would be a more rapid and perhaps even a more accurate classifier performance. Table A.5 of appendix A shows the full analysis of the backwards stepwise removal procedure, giving the values of Λ, F and λ for each feature as the feature set was reduced in number from 16 features to only one (which was $\overline{R}$). These three statistics are also graphed in figure 6.6

This graph shows that, as features were removed from the set, the slight degradation in classifier performance gave rise at first to a gradual increase in the values of Λ and F (because the within-groups variances were rising in relation to the overall variance of the training data for all of the features taken as a set). There was also a steady reduction in the partial λ for each successive feature removed, which meant that the process was removing features each of whose value to the discriminatory power of the set was slightly higher than the last. Eventually, the value of Λ began to rise very fast as the feature set became very small and its ability to separate the data into the correct classes began to collapse. The partial lambdas of the last few features to

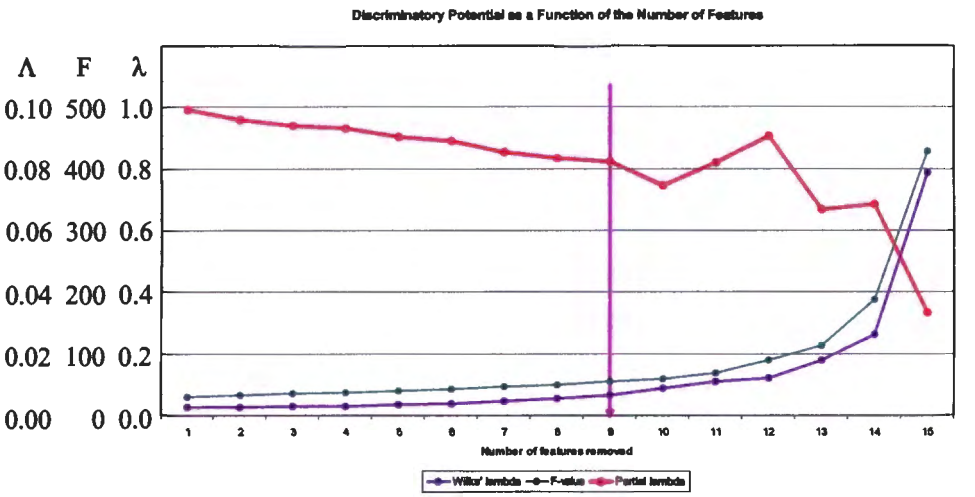


Figure 6.6: **Variation in Λ, *F* and λ for varying numbers of features**

be removed exhibited erratic variations in value, as the overall discriminatory power of the remaining feature set became more sensitive to the discriminating value of its individual members, or of small subsets of features within it.

Recognising that the reduction of a feature set will always involve some arbitrary decisions, it was decided to reduce the feature set in the colour classifier to the point indicated by the purple arrow, which was felt to be positioned safely before the onset of the instabilities mentioned above. This implied the removal of nine features and the retention of seven, which are listed in table A.6 of appendix A. These remaining seven features included none of the modal and standard deviation statistics, but they did comprise several of the skewness and kurtosis measures, which to some extent vindicated the approach of characterising the colour and intensity distributions by their higher-order moments. The raggedness feature was ejected quite early on (see table 6.4 and table A.5), which suggests that the low correlations which it exhibited in figure 6.3 were due more to its being a "noisy" statistic than to its adding fresh information. For the remaining seven features, table A.6 gives an impression of their relative individual values by listing the overall Wilks' lambdas that would result upon removal of each, their individual partial lambdas (which are all below the arbitrary value of 0.75), and confidence statistics (high F and very low p) which indicate for each feature that its removal from the feature set would be detrimental to the set's discriminatory potential at a very high level of statistical significance.

Under the operation of the reduced feature set, the training data formed clusters in 7-dimensional feature space. The squared Mahalanobis distances between the centroids of the five colour-class clusters are given in appendix A in table A.7. Knowing these distances assisted later in the interpretation of the test results. The final stage in the training of the colour classifier was the calculation of decision functions which would define the class boundaries in the 7-dimensional feature space. These functions were calculated by the Statistica package on the basis of the analysis of variance (ANOVA) of the clustering of the training data. The output of the package in this case was five functions in the form of weighted sums of the remaining seven features, plus a constant. The calculated weights and three possible values for the constant are given in table A.8 in appendix A. As an example, the decision function for the pale lemon class was

$$\begin{aligned} d_{ple} = {} & 1.3578\overline{R} + 1.9113 R_{\text{kurt}} + 0.3746\overline{G} + 70.5226 G_{\text{skew}} \\ & - 10.2235 G_{\text{kurt}} - 32.2419 I_{\text{skew}} + 6.7815 I_{\text{kurt}} - 137.9386 \end{aligned}$$

and that for the lemon class was

$$\begin{aligned} d_{lem} = {} & 1.7197\overline{R} + 1.4127 R_{\text{kurt}} - 0.4478\overline{G} + 33.8830 G_{\text{skew}} \\ & - 4.7444 G_{\text{kurt}} + 5.5353 I_{\text{skew}} + 0.8874 I_{\text{kurt}} - 102.0708 \end{aligned}$$

with the hypersurface dividing the two classes being given by the equation

$$d_{lem} - d_{ple} = 0$$

The constants in these decision functions are slightly sensitive to the distribution of the training data among the classes, as expressed by the *a priori* probabilities for each class. Table A.8 gives three possible values for the constant: the first is for use in defining class boundaries under the assumption that the class proportions are the same as those of the training data; the second is of value where nothing is known about the data class proportions either in the training set or in the "real world", and so they are assumed to be equal; and the third is used to define class boundaries for test data which has been drawn randomly from the "real world", and so is assumed to have the same class proportions as exist in the wider context. The "real world" colour class proportions for tobacco leaves will be discussed in the next section. The derivation of the decision functions concluded the design and training of the colour classifier: it now remained to test its performance.

6.4 Testing the colour classifier

The 100 leaf data samples that had been set aside during the training of the colour classifier were now classified with the use of Statistica, which first calculated the squared Mahalanobis distance of each test sample's feature vector from the class centroids. These squared distances are tabulated as the first set of columns in table A.9 of appendix A. The *a posteriori* probability that the sample belonged to each of the five colour classes was then calculated on the basis of these squared Mahalanobis distances. The five probabilities for each sample are given in the second set of columns of table A.9. Classification of the sample was then done on the basis of the maximum *a posteriori* probability, and the final columns of table A.9 give the order of likelihood in which the classes were assigned to the sample, with the most probable class listed first. An asterisk placed before the name of the sample in the first column of the table indicates that this classification turned out to be incorrect when compared to the label that had been given to the sample by expert graders.

Actual class	% correct	Pale Lemon	Lemon	Orange	Lt Mahog	Dk Mahog	TOTAL
Pale Lemon	83.33	10	2	0	0	0	12
Lemon	87.50	3	21	0	0	0	24
Orange	77.27	0	1	17	4	0	22
Lt Mahog	91.30	0	0	0	21	2	23
Dk Mahog	100.00	0	0	0	0	19	19
TOTAL	**88.0%**	13	24	17	25	21	100

Table 6.5: **Classification matrix for the unadjusted 7-feature colour classifier**

A summary of all of the classifications that were made in the test is given in the *classification matrix* shown in table 6.5. The table shows, for each class of the test samples, how those samples were allocated to classes by the classifier. So, for instance, of the twelve pale lemon leaves in the test data, ten were assigned to the pale lemon class, and only two were mistakenly graded as lemon, giving a success rate of about 83.3% with that class of tobacco. Overall, the classifier correctly classified 88 of the 100 leaves.

All of the calculations involved in producing table A.9 and table 6.5 were done under the assumption that the test data would conform to the same *a priori* class probabilities as the data in the training set. These probabilities, which appear in table A.9, were as

follows:

$$P_{ple} = 0.12; \quad P_{lem} = 0.24; \quad P_{ora} = 0.20; \quad P_{lma} = 0.24; \quad P_{dma} = 0.20$$

As was shown in the previous chapter, if the classifier is intended for use with "real world" data, then it is more appropriate to train, test and use the classifier under the assumption that data will be distributed according to the class proportions to be found in reality. An effort was therefore made to find the appropriate proportions for usage in this project.

Tobacco colour class percentages fluctuate quite widely in Zimbabwe, as elsewhere in the world, partly as a response to demand (which is fashion-driven and also dependent upon the types and quantities of tobacco stock held throughout the world), and partly as an unavoidable function of weather conditions, which will to some extent influence which styles of tobacco can successfully be grown. To obtain an impression of the "real world" class properties, the class proportions from Zimbabwean auctions over a seven year period were tabulated and averaged, as shown in table 6.6 below [81, 83]:

Year	Pale Lemon	Lemon	Orange	Lt Mahog	Dk Mahog	Unclassified
1990	0.61	46.04	32.46	3.72	0.93	16.24
1991	0.50	37.21	36.08	6.86	1.71	17.64
1992	0.64	47.95	29.03	3.39	0.85	18.14
1993	0.54	40.40	32.79	5.53	1.38	19.36
1994	0.62	46.29	34.56	4.76	1.19	12.58
1995	0.71	53.21	30.29	2.58	0.65	12.56
1996	0.53	39.89	39.50	4.27	1.07	14.74
Mean	0.59	44.43	33.53	4.44	1.11	15.90
%	0.7	52.8	39.9	5.3	1.3	100%

Table 6.6: **Proportions of the five colour classes averaged over seven years**

The final percentage figures are based on all classified tobacco, and so are calculated without counting the 15.9% of the total which is listed as unclassified. The mahogany proportions are approximate, with light mahogany assumed to be about four times as common as the darker style. One may note from this table both that the great majority of classified leaf tobacco is either lemon or orange, and that the fluctuation in the proportions of these grades from year to year are often far larger, as a percentage of the total, than the totals for the three minority grades. This has important implications

for a machine vision tobacco colour classifier: firstly, it should not be trained on the assumption of equal, or nearly equal, class proportions; and secondly, while it must be trained to recognise a minority class such as pale lemon, any errors made in classifying pale lemon leaves will, as a percentage of all of the leaves being classified, represent a very minor aberration. Conversely, the fact that the classifier incorrectly classified about 23% of orange leaves (see table 6.5) is a serious flaw in view of the fact that orange leaves contribute about 40% of the total number.

Table A.10 of appendix A gives the results that were obtained by training the 7-feature tobacco leaf colour classifier using the following modified *a priori* probabilities, which are typical of the actual colour class distribution:

$$P_{ple} = 0.01; \quad P_{lem} = 0.53; \quad P_{ora} = 0.40; \quad P_{lma} = 0.05; \quad P_{dma} = 0.01$$

Whilst the squared Mahalanobis distances were not changed by this modification, the *a posteriori* probabilities attached to the leaf sample classifications did alter. One can see, in comparing table A.10 with table A.9 that the modified classifier misclassified a slightly different combination of the 100 test samples than the combination that had been in error originally. The classification matrix that resulted from this second test is given in table 6.7, which records a substantially improved classification performance, especially in the important lemon and orange categories. Even in the minor colour classes, dark mahogany remained perfectly recognised and the performance on light mahogany was excellent. Only the performance on pale lemon leaves had deteriorated, and this was based upon only 12 test samples and, as has been mentioned, would affect a very minute percentage of the leaves in a practical application.

Actual class	% correct	Pale Lemon	Lemon	Orange	Lt Mahog	Dk Mahog	TOTAL
Pale Lemon	75.00	9	3	0	0	0	12
Lemon	95.83	1	23	0	0	0	24
Orange	90.91	0	1	20	1	0	22
Lt Mahog	91.30	0	0	0	21	2	23
Dk Mahog	100.00	0	0	0	0	19	19
TOTAL	**92.0%**	10	27	20	22	21	100

Table 6.7: **Classification matrix for a "real-world" adjusted 7-feature colour classifier**

The 92% scored by the adjusted classifier is a percentage that is based only on the performance on the 100 samples of the test set (and could be recorded with less as-

sociated error if the test set had been larger). On the assumption that the accuracy of this result is adequate, however, it can be further refined, as described in chapter 5, to predict the true success rate of this classifier when working to classify "real world" data as:

$$\frac{75}{100} \times 0.007 + \frac{95.83}{100} \times 0.528 + \frac{90.91}{100} \times 0.399 + \frac{91.30}{100} \times 0.053 + \frac{100}{100} \times 0.013 = \mathbf{93.5\%}$$

This final percentage of 93.5% is the predicted success rate for the colour classifier in this project, which is the 7-feature classifier trained on 152 selected data cases under the assumption of "real world" *a priori* sample probabilities. This classifier's performance might be further improved by training it on a larger number of data cases or by modifying it for operation in a particular place or for use during a particular grading season, if in each of these cases a set of *a priori* class probabilities were accurately known.

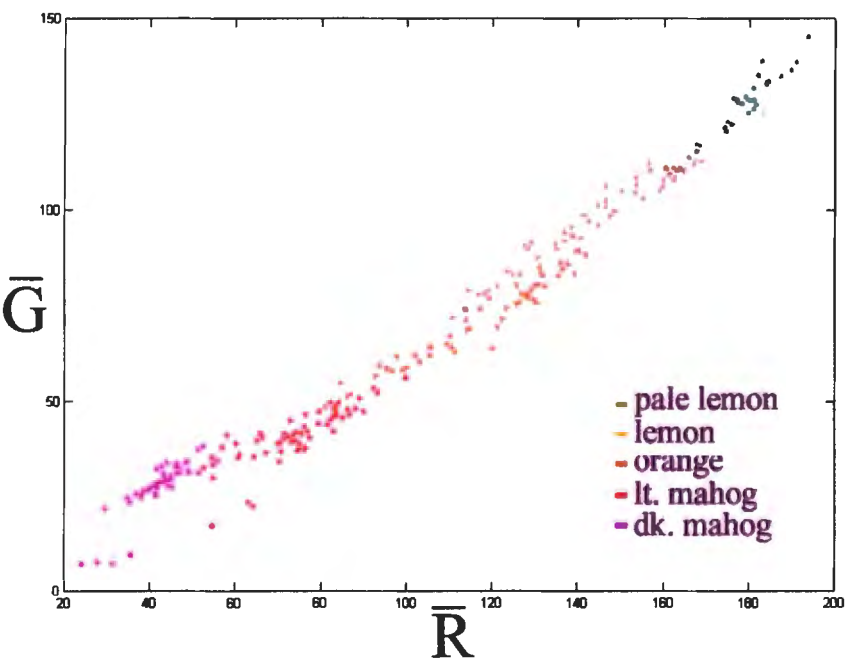


Figure 6.7: **Separation of 252 data points by only 2 features ($\overline{R}$ and $\overline{G}$)**

The speed of operation in an actual implementation of this sort of classifier would be determined in part by the computer hardware and software that was used. To speed up the *algorithm* still more would require a further reduction in the number of features. A maximum-speed version of the classifier, which required only two of the original

16 features ($\overline{R}$ and $\overline{G}$) was also investigated. This classifier performed very well, considering its simplicity, scoring 85% correct classifications under the assumption of a "real world" distribution, with a predicted performance of 92.6% on "real world" leaves. Its specific performance on pale lemon leaves was poor (58.3%), however, and the other minority colours were also not as well classified as by the 7-feature classifier, scoring 78% for light mahogany and 89% for dark mahogany. Nevertheless, a 2-feature classifier may be an attractive option where speed and a basic ability to distinguish lemon tobacco from orange tobacco is what is sought. Figure 6.7 shows all 252 data cases plotted in the 2-dimensional feature space of the fast classifier. The overlap between classes is evident, and it seems very possible that the classifier might perform more poorly on a different test set, but the efficiency of the two features $\overline{R}$ and $\overline{G}$ acting in concert remains impressive.

6.5 Archetypal leaves

Figure 6.8: **Five archetypal leaves, each representative of its colour class**

Besides producing a classifier to sort leaves into their colour classes, it was also a stated aim of this project to define the leaf classes quantitatively for the information of an industry which works largely on *qualitative* assessment of tobacco colour and by much handed-down experience. The statistical output of the classifier now makes it easy to define what is meant by a very typical (or *archetypal*) leaf of a particular colour class, and this might be useful information in the training of graders or for ensuring maximum consistency in persons who are already trained. Figure 6.8 shows five leaves which were selected from the test data on the basis that their *a posteriori* probabilities of class membership were the maximum for their class, and that they were one of the

closest leaves in their class to the class centroid in the squared Mahalanobis sense (see table 6.8). These are therefore leaves that were identified by the classifier as being extremely typical of their classes. Images of these archtypal leaves, or preferably the leaves themselves, would give a valuable impression to trainee graders of what is meant by each class colour. The images printed here will serve that purpose, despite some inevitable slight mutation of the colours during the printing process.

		Pale Lemon	Lemon	Orange	Lt Mahog	Dk Mahog
	Leaf data number	2e2	219	3o9	3r1	3s5
	A posteriori probability	1.000	1.000	1.000	1.000	1.000
	Squared Mahalanobis distance	6.05	2.51	3.64	1.62	1.69
	Number of pixels selected	269	339	343	216	298
R	Minimum	196	189	129	59	31
	Maximum	226	213	182	114	66
	Mean	209.46	199.96	154.84	89.32	47.26
	Std Dev	6.25	4.14	11.59	11.80	7.55
G	Minimum	144	128	70	25	16
	Maximum	182	165	120	67	42
	Mean	160.26	143.39	93.17	43.68	27.21
	Std Dev	6.99	5.73	10.50	8.66	5.17
B	Minimum	72	59	27	19	16
	Maximum	141	107	66	42	48
	Mean	92.27	73.78	43.03	27.85	30.74
	Std Dev	12.07	10.63	6.13	4.36	6.12

Table 6.8: **Measurements taken from sections of clear laminae of archetypal leaves**

Statistics for typical examples of the five colour classes may also be obtained by looking at the feature values for the leaves which are named as "2e2", "219", "3o9", "3r1" and "3s5" in table A.1 of appendix A. However, the mean values of R, G and I given for these leaves in the table could be misleading as indicators of the true background colour of archetypal leaves, because they are values taken from pixel information across the whole leaf, including many dark regions such as the butt, midrib, veins, and any spots or damaged sections on the leaf's surface. It is the *underlying* colour of the leaf lamina which is diagnostic of its true colour class; and to find an adequate definition for the typical unblemished lamina colours of the classes, the five archetypal

leaves of figure 6.8 were each studied further. Several regions of unblemished lamina were identified by eye on each of these leaves, and colour statistics were gathered from a sample of between 200 and 350 pixels in such undamaged areas. The R, G and B histogram measurements for these sampled pixels are summarised in table 6.8. Thus, for example, by rounding the mean values appropriately, it was found that the pale lemon class is typified by a lamina of colour $R = 209$, $G = 160$ and $B = 92$. The observed ranges and standard deviation values for each leaf colour are also given in table 6.8. These measurements give objective and repeatable definitions for the five colour classes of flue-cured Virginia tobacco. It may be noted that samples of background colour from unblemished regions of lamina would probably make excellent features for a tobacco leaf colour classifier. However, the segmentation of sometimes tiny sections of unblemished lamina is a very difficult task, even for a human being, and would be unlikely to be amenable to any straightforward image processing technique.

The information in table 6.8 was finally used to synthesise some swatches of colour that might be of value in colour grading. These swatches, which are shown together in figure 6.9, have very similar colour means and standard deviations to those stated in the table, and have been given a texture which mimics the surface of a tobacco leaf. Whilst their colour reproduction on the printed page is a possible source of error to which further attention should be given, swatches such as these would offer a grader a standard set of colours by which to make more accurate and consistent grading decisions.

Figure 6.9: **Swatches illustrating the underlying lamina colours of the five classes**

Chapter 7

Tobacco Leaf Plant Position Classification

7.1 Visual indicators of plant position

In the design of a classifier that would be able to grade leaf images on the basis of the leaf's *plant position* on the stalk at reaping, care was taken to employ features which, as far as possible, corresponded to the actual criteria that human graders use in making the same decision. As mentioned in chapter 2, several important considerations, such as the thickness and oiliness to the touch of the lamina, or such as the leaf's aroma, are non-visual and therefore not available to a machine vision classifier that operates only on leaf images. This reduction in the number of available cues makes an already-difficult task into a very challenging one; but it was nevertheless of interest to investigate how well an automated plant position classifier might perform using visual grading criteria only.

The visual considerations applied by graders in deciding from where on a tobacco plant a leaf has been reaped fall into three loose categories. Firstly, the overall size and simple dimensions of a leaf are good plant position indicators: primings are small, short, wide leaves, whilst lugs, cutters and leaf are progressively larger, longer and thinner as one looks further up the plant. High on the stalk, the smokers are once again shorter, but are also thin; and the tips are often the smallest, thinnest leaves to be reaped from the plant. Appendix C gives the pictures of all 210 leaves which were used in the development of a plant position classifier, and these simple variations in

leaf dimensions are quite evident in browsing through these images. So, too, are the considerable variations in shape and size *within* plant position categories, which is a difficulty with which a machine vision classifier must contend. Furthermore, damage to the outline and laminæ of leaves makes measurement of their size dimensions even more problematic. A human grader subconsciously visually interpolates a leaf outline where there is damage, and can usually assess the "size" of a leaf, even though parts of it are not there!

The second approach which human graders take towards identifying a leaf's plant position is an appraisal of the leaf's general shape by reference to their experience of the typical shapes of each class. It must be emphasised that graders have hitherto had very few quantitative yardsticks for this way of grading, but that many graders will confidently grade a leaf on the basis that it "looks like a typical lug", for example. The justification for this sort of approach is that lower leaves tend to be somewhat spatulate in shape, rounded near the butt and bulging towards their rounded ends, whilst the higher leaves are more lenticular and can be identified by their very angularly-pointed tips. Taking their cues from this, and from other more subtle and probably unconscious considerations such as the position of the concavities in the leaf's outline, experienced human graders will usually correctly identify a leaf's plant position. However, since they would invariably have had full access to touching and smelling the leaf, they would be making the grading decision on the basis of an overall impression based upon numerous indications, many of which they might find difficult to put into words. In an informal test for this project, one experienced farmer was asked to grade the leaves in appendix C on the basis only of their printed outlines, which were given to him in random order. Without the assistance of touching or smelling the leaves, he was able to identify the plant position group correctly in about 75% of the cases.

The third visual indicators of leaf plant position come from inspecting the veinous system of the leaf, especially the thicknesses of the stalklet or *butt* of the leaf and of the main vein supplying the leaf up its central line, which is called the *midrib*. In lower leaves such as primings and lugs, the butt and the midrib are both quite thin, and the midrib is not especially noticeable as a visible feature of the leaf. Further up the plant, the butt of the leaf becomes thicker, and the midrib longer and more prominent. An experienced grader would certainly use these as clues to the leaf's origin on the plant; but it will be seen from appendix C that visually appraising the midrib of a leaf is made quite difficult if the leaf colour is rather dark. In fact, there

is a tendency for leaves higher on the plant to be darker in colour after curing, and for lower leaves to be lighter; but this tendency was not exploited in this project since there are many counter-examples in actual practice and because of the limited numbers of leaf data images that were available. Regardless of the leaf colour, extraction of the butt and midrib for use in deriving features for plant position determination proved to be possible with the use of appropriate image processing techniques.

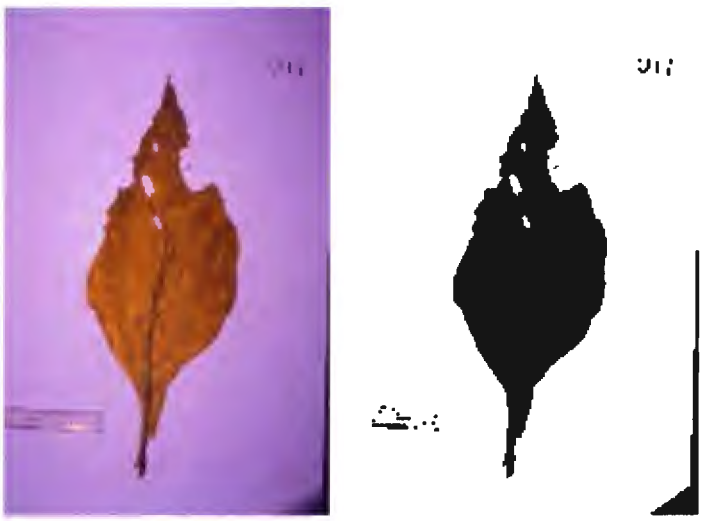
The next three sections of this chapter describe in detail how image processing algorithms were developed and used to capture as many as possible of the features by which it was believed that a human grader classifies tobacco leaf plant position. The three sections cover the three broad approaches to visual appraisal — size, shape similarity and veinous system scrutiny — that have been introduced above.

7.2 Reconstructing and measuring leaves

Before measuring the basic dimensions of a leaf (such as its area, length and width for use as features in the classifier) it was considered important to pre-process all of the leaf images so as to ensure that these measurements would not be unduly affected by damage either to the outline shape or to the interior lamina of the leaf being measured. Thus, the aim was to reconstruct the unbroken outline of a leaf by interpolating its outline contour in sections of damage, much as the human eye-brain system seems capable of doing when making an overall assessment of a leaf's size.

Because size measurements with an accuracy more precise than 1% were not considered necessary, and in order greatly to speed up the operation of the image processing algorithms, each 1700×2600 24-bit colour image that was to be used was first reduced to 147×225 pixels, to give an image such as the one shown in figure 7.1(a). The standard procedure of thresholding this image in the blue (B) band at $B = 170$ produced the binary image of figure 7.1(b), following which region-labelling and identification of the largest foreground object in the image yielded the segmented leaf object, given in figure 7.1(c). The removal of interior damage was achieved by once again region-labelling the binary image, identifying the "background object", and then inverting the image so as to arrive at the "non-background object", which is a representation of the leaf without any holes, as shown in figure 7.1(d).

For leaf size and shape analysis, what was required was an image of the leaf lam-



(a) Resample (b) Threshold

(c) Isolate (d) Fill holes

Figure 7.1: **Early stages in the reconstruction of a leaf**

ina region only, with an interpolated outline in sections of damage, and with the butt removed so that it would not influence measurements of length or the calculated position of the leaf's centroid, for example. The strategy of reconstructing the leaf in order to produce an estimate of the undamaged outline from contours such as figure 7.1(d) continued with the removal of the butt of the leaf by eroding the binary image with a circular disk of radius 40 pixels, so as to obtain the image of figure 7.2(a). The vertical co-ordinate of the lowest object (black) pixel in this eroded image was then

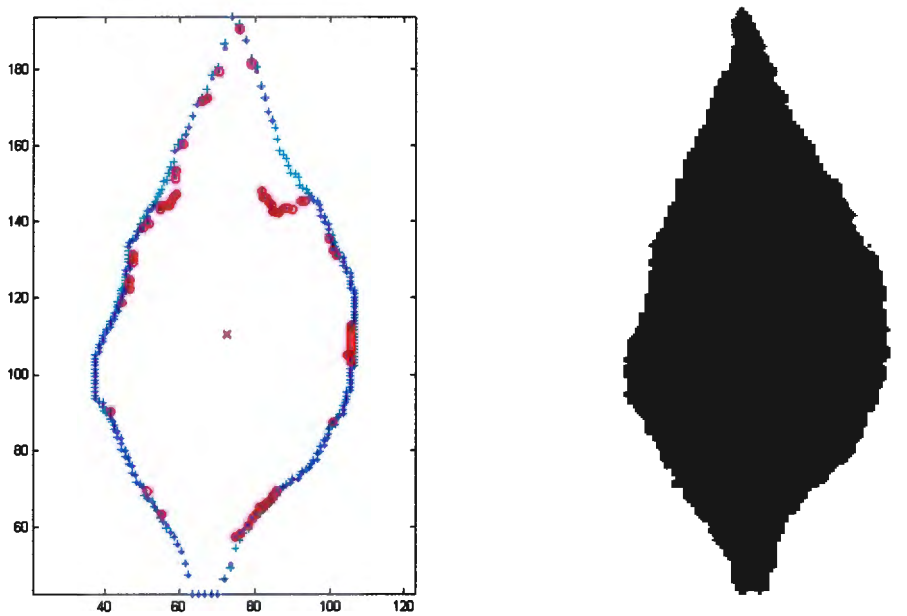
(a) Erode outline (b) Remove butt

(c) Closed image (d) Reveal Damage

Figure 7.2: **Continuing stages in the reconstruction of a leaf**

taken as the bottom point of the leaf, with all lower points being considered to be part of the butt. This division is somewhat arbitrary and is confused by the small flaps of leaf, called *petioles* on either side of the upper butt, but it was considered to give a reasonable segmentation of the measurable leaf area, and was applied identically to all of the leaf images so as to give a fair basis for comparison. Figure 7.2(b) shows the binary leaf object with its butt removed in this way. To assess which parts of the leaf outline had probably suffered damage, this image was now closed (dilated and

then eroded) using the same structuring disc of radius 40 pixels. The result is shown in figure 7.2(c), and is also shown with the binary leaf outline superimposed upon it in figure 7.2(d). Regions of outline damage are clearly seen in this superimposed view as places in which the silhouette of the closed leaf (in red) differs from that of the original leaf (in green).



(a) Outlines, with damage (b) Reconstructed leaf

Figure 7.3: **Identification of damaged lamina outline, and a reconstructed leaf**

The reconstruction of an undamaged leaf outline might adequately have been achieved by adopting the closed leaf outline of figure 7.2(c) wherever there was damage; but since the leaf shape was to be characterised partly in terms of its Fourier descriptors in any case, it was decided to use an elegant outline interpolation method that the use of the Fourier analysis of the leaf outline would make possible. Both the damaged outline and the closed leaf outline of figure 7.2(d) were sampled at 256 sample points, located around their boundaries at angles that were equally separated when subtended at the centroid of the damaged leaf. The number of samples was chosen to be large enough that the boundary outline could easily be reconstructed to good accuracy from a reasonably evenly-distributed half or third of the sample points, it having been shown

that 64 samples were sufficient to reconstruct all visible features very adequately, and was chosen to be a power of 2 so that the fast Fourier transform would run efficiently on the sample series. Figure 7.3(a) shows in dark blue the sample points that were taken around the original damaged outline, and also gives in light blue the sample points from around the closed leaf outline. Where the positions of two corresponding samples differed by a distance of more than one pixel, this was interpreted as an angle from the centroid at which outline damage existed. These positions of damage are shown in figure 7.3(a) with red symbols.

The sample series was next stored as an array of elements, g_r $(r = 1,2...256)$, from which every element which corresponded to a position of damage on the leaf outline (the red symbols in figure 7.3(a)) was *removed*, so as to leave a series of unequally-spaced samples, each a complex number representing the position of an outline point in 2 dimensions. The discrete Fourier transform of g_r was then computed, followed by the inverse Fourier transform, which returned a boundary function in which the missing outline sections were replaced by interpolated values that were based on the shape features of the entire outline. It was believed that this method of interpolation provided a reconstructed leaf outline which was very similar to the imagined leaf outline that the human eye-brain system would create in visually assessing the overall size and shape of a damaged leaf. Figure 7.3(b) shows this reconstructed leaf shape for the case of the example described above: the leaf object has also been uprighted by finding its angle with the vertical, θ, through the use of the Hotelling transform, and then by multiplying all of the Fourier descriptors of its outline by $e^{j\theta}$ prior to the application of the inverse Fourier transform. The dc Fourier descriptor, G_0, was also previously set to the position of the central point of the image so as to centre the leaf object in the frame of view after the inverse transform. Thus, the use of the Fourier transform interpolation method allowed the uprighting and centring of the image to be performed in a very convenient way. The Fourier descriptors of the reconstructed leaf outlines were now calculated from a set of 256 evenly-spaced samples for use as shape features, as will be discussed in the next section.

The isolation of the uprighted and butt-less leaf object of figure 7.3(b) now made it very easy to extract several size features which were considered likely to have discriminatory power in the plant position classifier. The distance between the object's maximum and minimum vertical co-ordinates was taken as its *absolute length*, and the difference between its maximum and minimum horizontal co-ordinates gave the

object's *absolute width* in a similar way. Both of these measurements were in units of *pixels*, which would be comparable across all of the images because they had all been acquired in an identical fashion and then all reduced to the same 147×225 resolution. The total number of pixels in the reconstructed leaf object gave the *area* of the leaf. To derive a feature that might distinguish between the lowest plant positions where the bulge of the leaf is closest to the stalk end and the higher plant positions where the maximum width of the leaf occurs progressively further towards the tip, the ratio of the length of the leaf to the vertical distance of the widest leaf width from the stalk end, $\frac{length}{height\ of\ widest\ width}$, was calculated. This is illustrated in figure 7.4, along with several of the other size features described here.

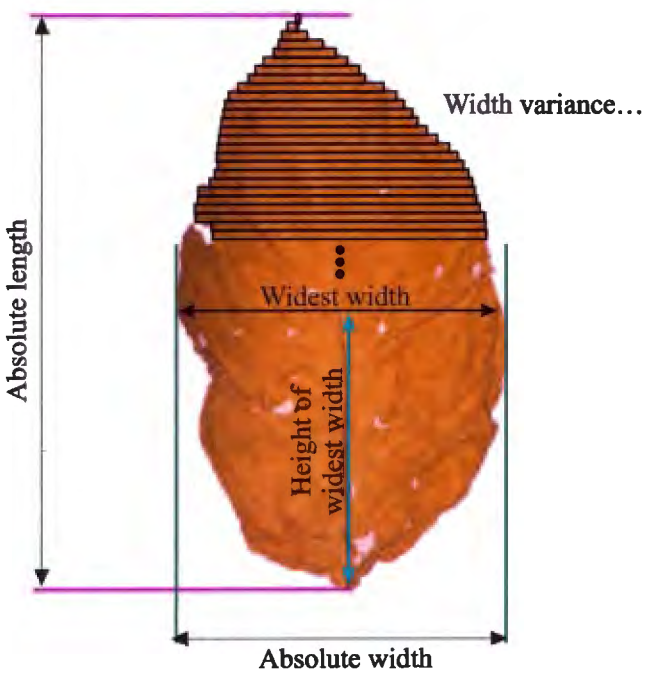


Figure 7.4: **Size features derived for use in the plant position classifier**

It was also considered likely that some impression of the outline shape of a leaf could be gained from measurements of how the width varied up the leaf, with width remaining much more constant as a function of height in the more straight-sided smokers or tips than it does in a spatulate and tapered leaf such as a priming or a lug. To indicate this variability for each leaf, the width was measured at every vertical co-ordinate value within the leaf, and the *width variance* was calculated from these measurements, as suggested by figure 7.4. Finally, to distinguish between short, wide leaves and long

tapered ones, the ratio of the absolute length to the absolute width, known as the *aspect ratio*, was derived as a sixth size feature.

7.3 Shape similarity and archetypal outlines

In the reconstruction of each leaf, the Fourier descriptors of the undamaged part of its sampled outline were initially calculated for the purpose of interpolating the outline and positioning and uprighting the leaf. The recalculated Fourier descriptors (G_u) carry information about the shape of the reconstructed leaf outline function (g_r) from which they were derived, with the lower-order descriptors, $G_{\pm 2}, G_{\pm 3} \ldots$ giving the gross shape whilst higher-order descriptors carry ever-finer detail about the outline. Bearing in mind human graders' propensity to assign a leaf to a plant position on the loose basis of its overall shape, it was decided to try to use the lower-order Fourier descriptors of each leaf as features in the plant position classifier.

Since, as figure 3.24 in chapter 3 shows, the shape of a leaf outline is well-captured by using between 7 and 10 Fourier descriptor pairs, it was decided to work with the 16 descriptors $G_{\pm 2}, G_{\pm 3}, \ldots G_{\pm 9}$ only. G_0 was not used, since it carries no shape information, and $G_{\pm 1}$, which defines an ellipse whose area is close to proportional to the area of the leaf object, was also not used since it would be redundant in conjunction with the size features derived in the previous section. Instead, each of the Fourier descriptors $G_{\pm 2}, G_{\pm 3}, \ldots G_{\pm 9}$ was divided by G_1 so as to bring every leaf object to a standard size. The resulting ratios, $G'_{\pm 2}, G'_{\pm 3}, \ldots G'_{\pm 9}$ then carried only shape information and could conveniently be used as features in the classifier. The inverse Fourier transform of the full set of descriptors (each divided by G_1) provided for each leaf a reconstructed boundary that was centrally positioned, scaled to standard size, and which was still sampled at 256 points. It was now possible, by averaging the values of corresponding sample points across many leaves in each class, to arrive at a set of 256 samples which defined the *mean leaf outline* for each of the six plant position classes. Figure 7.5 shows the outline of the mean priming that was calculated in this way.

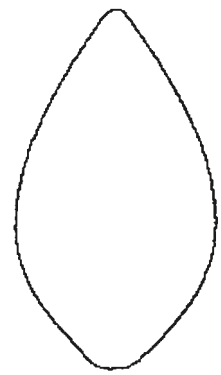


Figure 7.5: **Mean priming**

It will be observed that because of the reconstruction process and because of the averaging, this outline is very smooth, and it seems to give an excellent indication of the typical shape of a priming. Because it was derived from 30 leaves whose tips varied in both position and orientation (and were sometimes not even present — see figure C.1), the mean priming is unrealistic as a tobacco leaf in the one sense that it has a rounded end instead of a tip. In terms of the Fourier descriptors, its tip is rounded because the tip of each individual leaf represents a sudden change of direction in the leaf's boundary and hence tends to contribute to the higher-frequency information that is carried by the higher-order Fourier descriptors. Since the mean leaf is an averaging of leaf outlines that were each reconstructed from a finite set of Fourier descriptors, the fine structure of the tips of each leaf was not well-preserved.

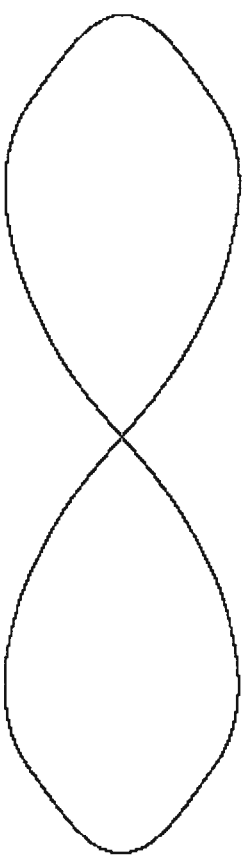


Figure 7.6: **Boundary function**

In order to create a realistic mean outline for each plant position class, every leaf was represented by a new sampled boundary function which consisted of two copies of the outline, end-on-end as shown in figure 7.6. Samples were taken around this boundary in the sense of a figure-of-eight, so that the new boundary function had 512 sample points and twice the period of the original leaf boundary function, g_r. It was found that the number of sample points around the boundary could be reduced from 512 to 32 (by taking every sixteenth point) without visibly reducing the accuracy of leaf reconstruction via the inverse Fourier transform, and that this reduction also had the benefit of reducing the number of sample points that were taken in the region of the leaf tips, where positional variance was at its highest. By sampling the figure-of-eight outline for 30 leaves in a class at 512 points, reconstructing the outline from damaged sections as described above, down-sampling to 32 sample points and then averaging across all 30 leaves, a mean outline for (half of) the figure-of-eight function gave an excellent and realistic impression of a typical leaf. Figure 7.7 shows in red the outline of the mean priming calculated in this way, superimposed on the original mean priming with

rounded tip, shown in blue. Except for the region around the tip, the correspondence is almost perfect despite the down-sampling to 32 sample points.

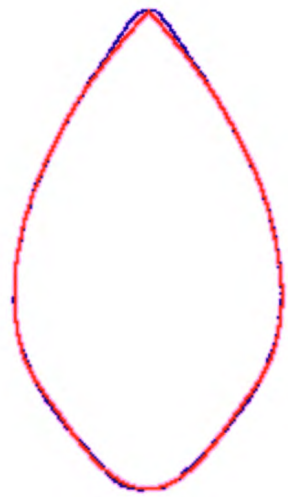
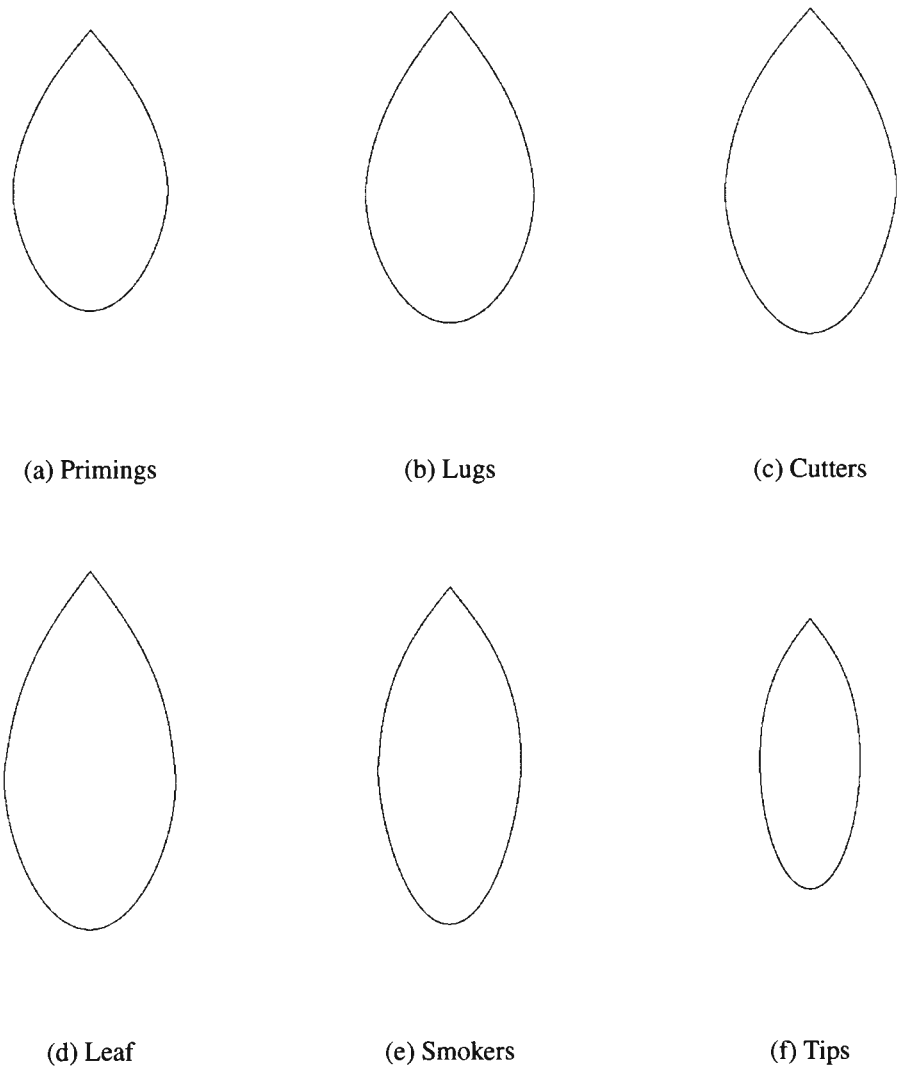


Figure 7.7: **Mean priming with reconstructed tip**

The averaged leaf outlines for each of the plant position groups were then calculated using these techniques, and the results are shown in figure 7.8. These archetypal boundaries represent an objective statement of what are meant by the shapes of the six plant position categories in flue-cured Virginia tobacco. Furthermore, because the outlines from which they were derived were *not* normalised to standard size through the division of the Fourier descriptors by G_1, they also give a good impression of the relative sizes of the different plant position groups. As such, they could be of value to graders as an objective expression of the size and shape variation of the leaves at all levels of a tobacco plant.

For the purposes of this project, the mean outlines (without the artificial pointed tips) also offered a way of deriving further shape features for use in the plant position classifier. When each leaf for classification had been boundary sampled (following reconstruction, centering, uprighting and normalisation to standard size), its boundary sample positions could be compared to the corresponding positions on each of the mean leaf template outlines for the six plant position groups. The template matching



(a) Primings (b) Lugs (c) Cutters

(d) Leaf (e) Smokers (f) Tips

Figure 7.8: **Archetypal leaf outlines for each of the six plant positions**

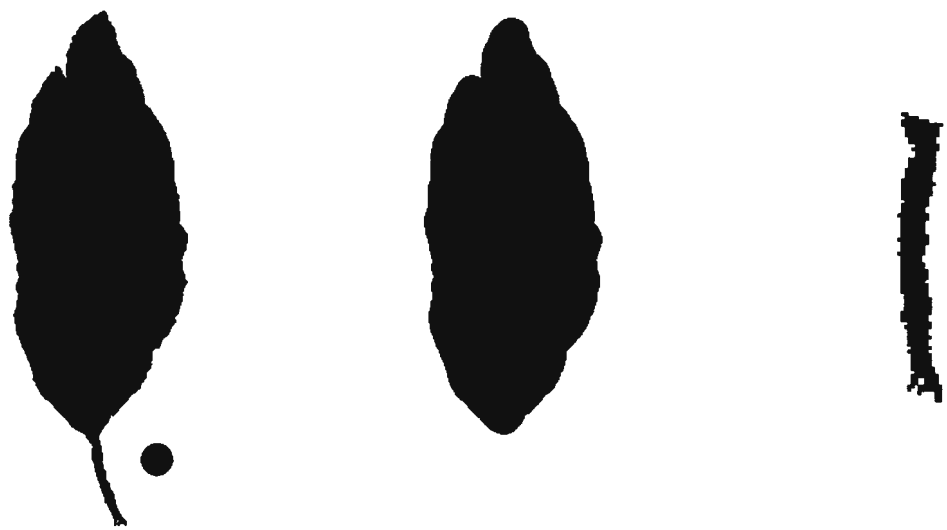
procedure is illustrated in figure 7.9. For each group, the sum of the squared errors of the sample positions gives a measure of the fit of the leaf to the mean template. This was expected to provide a powerful set of features, the lowest-valued of which would be strongly indicative that the leaf should be a member of the corresponding class. As a method of plant position diagnosis, this has its analogue in the approach of the human graders who, as discussed above, will often assign a leaf to a plant position group on the basis of its similarity to their general concept of typicality within the group. Using the six mean templates, the machine vision classifier has the advantage of being able to do this in an objective and consistent manner.

Figure 7.9: **A tobacco leaf, shown in comparison with the mean priming template**

Only in the region of the tip of the leaf is the template matching approach liable to discriminate poorly between classes, and yet the pointedness of the tip of a tobacco leaf is cited by many human graders as a good indicator of plant position. It was therefore decided to measure the angle of the pointed tip of each leaf, and to use that measurement as another shape feature in the classifier. The *tip angle* was calculated by considering the 16 sample points of the 256 outline samples which lay closest to the top of the leaf (which was assumed to be the tip). A straight line of best fit was constructed through those members of this group of 16 that included or lay to the left of the uppermost point, and another straight line of best fit was constructed for the subset of the 16 points that included the tip or lay to its right. The angle formed where these two straight lines crossed, calculated using the Cosine Rule discussed in chapter 3, was then denoted as the tip angle. Together with the 16 normalised Fourier descriptors and the six sums of squared errors to template fits, the tip angle concluded a set of 23 candidate shape features which were derived for each training leaf and considered for use in the classifier.

7.4 Butt and midrib measurements

The final features that were considered for automatic plant position classification were derived from measurements of either the *butt* (the lower stalk) of the leaf or else of its central vein or *midrib*. Because these objects are both much smaller than the entire leaf, it was felt appropriate to work with images of somewhat higher resolution than had been used to extract other features, and so the archived 1700×2600 pixel leaf images were each reduced to dimensions of 340×520 pixels for this purpose.



(a) Leaf & structuring element (b) Opened leaf (c) Upright difference image

Figure 7.10: **The segmentation of a leaf butt by binary opening**

It has already been seen (as in figure 7.2) how morphological erosion of a binary leaf image can be used to remove the protruding outer features of the leaf, including the butt. Although a disc of radius 40 pixels worked adequately as a structuring element in segmenting out the leaf object, for the specific segmentation of the butt a finer control on the size of the structuring element was required, especially as the size of the leaf object between images was prone to considerable variation (see appendix C). After some experimentation, it was found that opening (eroding and then dilating) using a disc with a radius of 7% of the square root of the number of pixels in the leaf object invariably did an excellent job of removing the butt of the leaf without unduly

eroding any of the much wider adjacent leaf lamina. Figure 7.10(a) shows a binary leaf silhouette which was then opened in this way to give the image of figure 7.10(b). The largest object in the image formed by taking the difference between the original image and the opened image was the butt, which is shown, enlarged and after alignment with the vertical axis by the use of the Hotelling transform, in figure 7.10(c).

The length of the butt is a haphazard function of where it was cut at reaping, and is of no value as a feature for plant position classification. However, the width of the butt was thought to be worthy of inclusion in the classifier, since it has some tendency to increase in moving from the thin leaves near to the ground to the fleshier, heavy leaves high on the plant. The *average butt width* was therefore calculated by taking the mean of the width in pixels at vertical co-ordinates down the length of the butt, excluding the top 25% and the bottom 25% to avoid any residual influence from either the petiole or the ragged end where the butt had been reaped from the stalk. This measure was then passed to the classifier design stage as another candidate feature.

Extracting the midrib of each leaf was a considerably more difficult process, especially as the midrib lay on the underside of the leaf lamina in all of the images and was always in danger of being obscured either by overlying flaps of the lamina if the leaf was not laid extremely flat, or by lack of contrast with darkened sections of damaged lamina. The best possible contrast and definition for the midrib was obtained by taking a grayscale image of each leaf, as in figure 7.11(a), and then applying histogram equalisation so as to arrive at an image in which the midrib was much better separated in intensity from the surrounding areas of bright lamina (see figure 7.11(b)).

(a) Grayscale leaf image (b) Histogram equalised

Figure 7.11: **Histogram equalisation of a grayscale leaf prior to midrib extraction**

It was next required to extract the long, vertical midrib preferentially from the surrounding blotchy regions of the lamina, and this was achieved using techniques in both grayscale and binary morphology. By closing (dilating and then eroding) the grayscale image of figure 7.11(b) with the grayscale structuring element whose intensity values are shown below

0	0	0	0	0	0	0	0	0
0	1	3	3	4	3	3	1	0
0	0	0	0	0	0	0	0	0

the image of figure 7.12(a) was formed. Remembering that the structuring element is of size 7×1 pixels and therefore very small in comparison with the whole leaf image, it is possible to visualise the grayscale closing as the passing of a "sledge", whose sledge-runner profile is given by the numbers in the structuring element, over a rocky terrain whose heights are given by the intensities in the leaf image. In this metaphor, the midrib represents a narrow gully over which the sledge smoothly passes, whilst the other parts of the leaf represent plateaux and basins to which the height of the much smaller sledge readily conforms as it passes over them. The butt, which is much wider than the midrib, is thus unaffected by the closing operation.

(a) Closed leaf image (b) Result of subtraction (c) Image after thresholding

Figure 7.12: **Early stages in the extraction of the midrib: subtraction of the closed leaf**

When the image of figure 7.11(b) was subtracted from that of figure 7.12(a) and then inverted, the result was the image shown in figure 7.12(b), in which the midrib was now becoming apparent as a series of elongated objects within this grayscale image. The butt and most features of the leaf lamina were no longer seen, and, apart from the midrib, there only remained a few vertically-aligned structures that had survived the morphological closing process. These are clearly visible in figure 7.12(c), which is a thresholded version of the previous image, thresholded at a grayscale value of 100 in order to return a binary image for the final extraction of the midrib.

The final identification of the midrib required several processing steps. The image of figure 7.12(c) was first binary dilated using the "dot" shown below as a structuring element.

0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

This had the effect of thickening every object in the image, including the unwanted structures — the result is shown in figure 7.13(a). Next, the image was eroded using as a structuring element the vertical "bar" which is shown below.

0	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	0

Long vertical objects such as the midrib were further enhanced by this process whereas small and horizontal objects were not favoured (see figure 7.13(b)). At this point, the image was region-labelled and its largest object was identified: it was assumed that this object, shown in bright blue in figure 7.13(c), formed part of the midrib. Then, as illustrated in figure 7.13(c), all objects that were less than 40 pixels in area (shown

(a) Dilate with "dot" (b) Erode with "bar"

(c) Apply removal rules (d) Midrib segmented

Figure 7.13: **Later stages in the extraction of the midrib: binary morphology**

in pink) were removed (by replacing them with the background colour), and the few surviving objects were tested and also removed either if their major axis deviated by more than 10° from the vertical or if the straight line between their centroid and the largest object's centroid made an angle of more than 10° with the vertical (these are shown in red). It now remained to restore the surviving objects, which in this case are the three sections of midrib shown in blue, to their original dimensions, and this was done by using the vertical "bar" to dilate the image and then by using the "dot"

structuring element to erode it. The outcome of the whole procedure was the segmented midrib, very slightly smoothed after the various morphological operations but still, it was believed, close enough to its original form to be of value for the extraction of useful features for the plant position classifier. The segmented midrib is shown, superimposed on the leaf outline, in figure 7.13(d).

The *average midrib width* was calculated for each leaf, using along the whole midrib length a similar algorithm to the one that had been developed to find average butt width. The total number of pixels in the midrib was also identified as a measure of its prominence, and so the *midrib area* was adopted as a possible feature. Finally, in order to express the midrib's size and prominence in relation to the overall leaf size, the two ratios $\frac{midrib\ length}{leaf\ absolute\ length}$ and $\frac{midrib\ area}{leaf\ lamina\ area}$ were added as candidate plant position features. Together with the mean butt width feature, this brought the number of features extracted from the veinous system of the leaf to five, and the total number of candidate features to 34.

7.5 Feature reduction and classifier results

The design of the plant position classifier now proceeded through the establishment of an effective discriminatory set of features for the model base and then on to the training of the model. Of the leaf images shown in appendix C, 30 were selected at random from each plant position class to be used as training data and the remaining 30 leaves (5 from each class) were retained for later use as test data. The 180 training data leaves were submitted to batch processing in order to derive values for all 34 candidate features for every leaf. The full table of this information was then submitted to the discriminant analysis package of Statistica, using a forward stepwise algorithm to select from the 34 candidates a set of features that would have good discriminatory power in plant position classification. The ten features that were selected for final inclusion in the classifier on the basis of this analysis are listed, in the order in which they were chosen, as the components of the feature vector shown overleaf.

$$\mathbf{x} = \begin{bmatrix} \text{width variance} \\ \text{leaf area} \\ \text{error in fit to mean smoker template} \\ \text{tip angle} \\ \text{absolute length} \\ \text{error in fit to mean tip template} \\ \text{absolute width} \\ \text{error in fit to mean leaf template} \\ \text{error in fit to mean priming template} \\ \text{error in fit to mean cutter template} \end{bmatrix}$$

Table B.1 of appendix B.1 summarises the forward stepwise analysis by which this feature set was obtained. At each step, a feature was added to the set on the basis of its having the highest *F-to-include* statistic: these F-statistics are given in column 3 of the table, followed in column 4 by the confidence statistic for the inclusion of the feature at the time of its addition to the feature set (a low value of p indicates a high confidence for the inclusion of the feature). The overall discriminatory power of the feature set is given by the calculated value of Wilks' lambda in column 5, and this can be seen to have improved (Λ falls) with each new feature added.

The feature selection process chose four of the leaf size features discussed in section 7.2, five of the overall leaf shape features described in section 7.3, and the tip angle. The use of the sixth leaf template (the squared error in fit to the mean lug template) was presumably redundant in view of the use of the other five. None of the Fourier descriptors proved to be a useful feature in this context. Also, none of the chosen set of 10 features was derived from the butt or midrib of the leaf's veinous system that was dealt with in section 7.4, and this may well have been because these features were "noisy". The midrib was extracted under difficult conditions, with the lamina sometimes obscuring it, and furthermore the morphological process of segmenting it may have significantly affected its measured width and area dimensions. The butt, too, is a relatively thin feature whose variation in width between the different plant positions may have been too subtle for image processing to capture, with fine shadows in the images sometimes adding slightly to the apparent butt width by way of further confusion.

The values to the classifier of each of the 10 selected feastures taken individually are summarised in table B.2 in appendix B.1. Given the initial discriminatory power for the feature set of 10 features as $\Lambda = 0.088117$, the table lists the Wilks' lambda for the remaining set of nine features that would result upon the removal of each of the 10, and the partial lambdas that can therefore be derived for each feature. An F-to-remove statistic and an associated confidence level are also given. Since the adverse effect upon the classifier of removing any feature seemed to be about equal for every feature, it was concluded that the chosen set of 10 features would not be improved in discriminatory power by further reducing it.

Discriminant analysis using these 10 features and the 180 leaf images of the training data was now performed. Table B.3 of appendix B.1 gives the coefficients for the decision functions for each of the six plant position classes, from which decision surfaces that separate the class regions in the feature space may be derived as described in previous chapters. This concluded the training of the plant position classifier.

In testing the performance of the classifier, there were two main problems. The first was that even when the classifier was used to test the 180 data images upon which it had been trained, it correctly identified only about 69% of cases. This performance is summarised in the classification matrix of table 7.1.

Actual class	% correct	Primings	Lugs	Cutters	Leaf	Smokers	Tips	TOTAL
Primings	80.00	24	4	2	0	0	0	30
Lugs	56.67	6	17	4	2	1	0	30
Cutters	63.33	4	1	20	4	0	1	30
Leaf	70.00	0	3	2	21	4	0	30
Smokers	66.67	2	0	2	5	19	2	30
Tips	80.00	0	1	0	0	4	25	30
TOTAL	**69.4%**	36	26	30	32	28	28	180

Table 7.1: **Classification matrix for the plant position training set of 180 leaves**

The reasons why this classification rate was not higher can be intuited by looking once again at the leaf image data in appendix C, where the variability of leaf size and shape within classes is quite evident. Some of this variability was due to the wide size and shape ranges that do exist in the appearance of leaves, even though they have been reaped from the same positions on the plant; and some of it is felt to be due to the fact

that, despite all efforts to ensure accurate pre-grading, there nevertheless remained some leaves which were pre-assigned to an incorrect category. Errors of this sort are more understandable when one considers that the tobacco leaves used as data had been removed from graded *bales* in complete *hands*, from which one or two leaves might yet have been incorrectly classified. No attempt was made to "clean up" the data, because the existing discrepancies, even after several rounds of expert scrutiny, are themselves representative of a genuinely difficult classification problem. It should be stressed, however, that the number of pre-grading errors in the plant position classifier data was felt to be rather greater than had been the case for the data set used in colour classification.

The second problem was the relatively small size of the test set, which consisted of the first five leaves from each of the class sets in appendix C. The paucity of test data was unavoidable in view of the time taken to acquire the images, yet it was nevertheless believed that a test set of 30 images would at least give an indication of how well the classifier was working. The classification matrix for the test set is given in table 7.2.

Actual class	% correct	Primings	Lugs	Cutters	Leaf	Smokers	Tips	TOTAL
Primings	80	4	0	1	0	0	0	5
Lugs	80	1	4	0	0	0	0	5
Cutters	80	0	0	4	0	1	0	5
Leaf	60	0	2	0	3	0	0	5
Smokers	60	0	0	1	1	3	0	5
Tips	60	0	0	0	0	2	3	5
TOTAL	**70.0%**	5	6	6	4	6	3	30

Table 7.2: **Classification matrix for the 30 members of the plant position test set**

Although the small numbers in the test data make the uncertainties on the results rather high, it is nevertheless clear that the classifier was correctly grading about two-thirds or three-quarters of the leaves, which was about as well as the expert human had performed when asked to state the plant positions of the samples using only pictures of their printed outlines. The 70% scored by the classifier on the test data is a performance that would very likely improve if the classifier could be trained on a larger and more consistent training set. In addition, a more precise impression of the performance could be had if the test set were larger.

The complete results from running the plant position classifier on all of the available data are tabulated in table B.4 of appendix B.1. This table gives the squared Mahalanobis distances of each leaf's feature vector from the class centroids in the 10-dimensional feature space, and also lists, in the second set of columns, the *a posteriori* probabilities which the classifier ascribed to the leaf's membership of each class. Classification was made, as in the colour classifier, on the basis of maximum *a posteriori* probability. Classification errors are shown by means of an asterisk in front of the leaf data number in the first column.

The full classification results allowed the most typical leaves in each class to be identified on the basis of minimum squared Mahalanobis distance from the class centroids. These leaves, one from each plant position class, data numbers 10, 44, 86, 122, 160 and 191, are depicted in figure 7.14. As typifiers of the six plant position classes, these leaves (or images of them) may be of value either in giving training or else in ensuring consistency in the grading of flue-cured Virginia tobacco, especially if they are used in conjunction with the archetypal leaf outlines of figure 7.8.

(a) Primings (No 10) (b) Lugs (No 44) (c) Cutters (No 86)

(d) Leaf (No 122) (e) Smokers (No 160) (f) Tips (No 191)

Figure 7.14: **Leaves closest to the class centroid for each of the six plant positions**

Chapter 8

Interpretation of Results, and Conclusions

8.1 Comments on the results

Before drawing the final conclusions of this project, some of the results should be qualified in the context of the practicalities of automating the tobacco grading process, and in terms of the way that data was acquired and handled in this investigation.

In the first instance, this project has *not* dealt with any of the problems that would be involved in the transport or mechanical handling of tobacco by any automated grading machine. There would be extremely daunting design issues in bringing tobacco leaves in bulk to the input side of such a system, moving them through the system below cameras, separating and suitably exposing individual leaves for data image capture, and then sending the leaves on to appropriate sorted lots, speedily and without causing any extra damage to them. Instead, this investigation worked with images of leaves that had been carefully laid flat, exposing the upper part of the lamina to full view. The conditioning and spreading of a leaf for data capture was a slow process, but without it there would have been no chance of extracting shape information suitable for use in the plant position classifier. In fact, the performance of the plant position classifier might well have been improved if it had *also* worked on images of the undersides of spread tobacco leaves, in which the midrib and the veinous system would have been better exposed to view. As far as colour classification is concerned, even though the underside of a leaf has a somewhat duller tone than the top surface of the lamina,

the results of this investigation do give some reason to hope that automatic colour classification of unspread leaves might be possible in a practical system based on algorithms similar to the ones developed here.

The second qualification of the results in this dissertation is that they apply to leaf image samples that were photographed under a particular set of geometrical and lighting conditions. These were made to be as consistent as possible, and were in accordance with industry specifications for grading illumination, and so they have given results that are coherent and repeatable, but which might have been different if the leaf images had been captured at a different distance or angle, or under different lights. A practical grading system based on the methods employed here would require to be calibrated during its commissioning so as to take proper account of the consistency, intensity and colour value of the lighting used, and of the geometry and optical characteristics of the imaging system.

Thirdly, little attention was paid to the speed of grading in the development of algorithms in this project. Certainly, every algorithm was implemented so as to run reasonably rapidly and efficiently within the context of this investigation, but the extraction of all of the features for the colour and plant position classification of a leaf still required of the order of a minute of computing time for each leaf sample. A human grader can handle and decide upon the grade of a leaf a good deal faster than that. Whilst there is promise that fast automatic grading methods (such as the 2-feature colour classifier described in chapter 6) could be developed, this was not a primary concern here.

Greater speed will, in general, be gained at the expense of grading accuracy, and, once again, the final judgement as to whether an automated system is "accurate enough" would probably have to be made by comparison with human performance. In an industry that is otherwise very rich in statistical information, figures for the grading accuracy of human tobacco classifiers is surprisingly scant. Farm graders, who are not usually working with the official grading scheme in mind, sort leaves into about a dozen piles: it is conceivable that they are correctly grading about 90% of the leaves, in the sense that a co-worker who re-graded the leaves might allocate about 10% of them to different piles (and yet the work of both graders would be acceptable to the foreman). Professional classifiers at the auction floors, on the other hand, work by the official grading scheme and are considered very consistent — although this does not appear to have been tested in a formal way, except insofar as there are apparently

very few complaints at auction which focus upon inaccurate grading by the classifier. An informal enquiry revealed that, while classifiers sometimes have difficulty with grading the more unusual grades and rarer styles of tobacco, their overall grading accuracy might be approximated as "90-95%". Automated colour classification may compete with this level of accuracy, but computer assessments of plant position and, one may speculate, of quality and of subtle differences of style, would probably fall short of expert human performance.

It will be noted that in the colour classifier results for this project and, to a lesser extent, in the plant position classifier results, errors in classification are generally committed by ascribing a leaf to an *adjacent* class. One is accustomed to thinking of an error within a classification scheme as being *absolute*, but in this context an error of only one class could be interpreted as a minor mistake in certain restricted cases. A lemon tobacco leaf graded as orange at auction would almost certainly by rejected by potential buyers, but a pale lemon tobacco leaf graded as lemon might receive more sympathy, simply because of the rarity of pale lemon tobacco and the debatability of where its boundary with true lemon colour really lies. Likewise, lugs put on auction as smoking leaf would probably receive very adverse comment, but lugs graded as primings or cutters might just be accepted. One is dealing with the classification of colours and shapes that are subject to *continuous* variation between individual leaves, and imposing upon them a *discrete* and artificial grading structure. Largely differing grades of colour or plant position will have different end usabilities in terms of their different leaf consistencies and chemical properties. Adjacent grades merge with each other by every measurable visible feature that was investigated in this project, and would presumably also do so in terms of other characteristics such as texture, aroma and chemical composition. Thus, the methods of this project, if used to grade tobacco by colour and plant position to a lower resolution (i.e. into a smaller number of grade categories), whilst never being 100% accurate might nonetheless be extremely effective in an application where small differences did not matter.

While every effort was made in this project to train the classifiers using a variety of leaf types to ensure robust classification of the unseen samples in the test sets, the training sets for an industrial system would need to be larger in number and still more diverse. Of particular concern in terms of the results of the plant position classifier was that the leaf samples were taken from only a few bales for each group. Better plant position classification would almost certainly be possible using a classifier trained on

many more than 180 images, taken from a much wider representative sampling of the leaf population. Clearly, this would have implications in terms of resources such as the time taken for image acquisition and preprocessing, or the space available for the storage of data, during the training phase.

There have been two valuable outcomes from this project apart from achieving its fundamental aim of developing machine vision classifiers for tobacco grading. The first of these has been the acquisition of a fine database of photographic images of tobacco leaves, which it is hoped may be of use in other research endeavours. Many of the images in the database are already digitised and stored on CD ROM for further computer-processing use. The second outcome has been the derivation of quantitative descriptors for the various colour and plant position tobacco grades that are in general use today. Insofar as these have been derived statistically from a wide range of representative leaf data, they should serve to inform the industry of what has been meant by each grade, by presenting parameters that are based upon numerous measurements so as to yield (for the first time) fully objective, consistent and repeatable colour and shape information for each grade. Further information, such as the mean lengths or areas of the leaves within each plant position group, was not explicitly calculated here, but is now easily extractable from the available image data. The typical colours and shapes of each class have been presented here in the form of *archetypes* and of images of typical leaves, because assessment of tobacco leaves is still very effectively done by comparison.

The last comments that should be made in interpreting the practical value of the results of this project have to do with social, financial and political issues. Automation of any process cannot be viewed in isolation to its alternative, namely that the work be done by human labour. Automation of tobacco grading might alleviate the lives of many people whose grading work is arduous and dull: on the other hand, it would also displace most of them from their employment, which is often their only hope of monetary income. In the wider context, it is tempting to characterise the tobacco industry as exploitative of cheap labour in the production of an unhealthy product; but this is a facile interpretation. Tobacco has been in extremely strong demand wherever it has been known for the past 500 years. Efforts on the supply side to farm it, to process it for consumption and to market it have certainly occupied the lives of many people over that time, but the market is nevertheless predominantly demand-driven. This much is clear from the addictive nature of tobacco smoking and from the fact that

such a large fraction of the retail prices of tobacco products is paid to governments as tax. Technology serves, as in all other products, to meet the demanded output; and it is employed to do so wherever it is economically viable and socially acceptable. In Zimbabwe, automation of farm grading would be most unlikely to be profitable for the forseeable future, because of the cheapness of agricultural labour. However, the results of this project may be of immediate use in other countries where a similar grading problem exists and where farm wages are much higher. The results may also be of value, in Zimbabwe and elsewhere, in the professional classification of baled tobacco at auction.

8.2 Conclusions

Algorithms have been developed which perform the classification of flue-cured Virginia tobacco leaves according to the standard grading scheme prevalent in Zimbabwe and elsewhere. The classifier derives seven features from a digitised image of the flattened spread leaf, and returns one of five colour classes for the leaf. The classifier may be expected to classify 93.5% of such leaves correctly, when the leaves are randomly selected from those available at auction.

An estimate of the typical lamina colour for each of the five colour classes has been derived, stated as a set of RGB values for each class and printed as a sample colour swatch. These results may be used as objective descriptors of the colour grades for the purposes of training graders or of developing grading consistency.

Further algorithms, which classify flue-cured Virginia tobacco leaves by the position on the plant stalk from which they were reaped, have also been developed. The plant position classifier measures ten size and shape features from a digitised image of a flattened spread leaf, and returns one of the six standard plant position classes. The classifier correctly classified 70% of cases in the test set, and misclassified leaves into an immediately adjacent class on a further 17% of occasions.

The leaf data processed in the development of the plant position classifier was also used to derive six archetypal shapes, which typify the shapes of the six plant position classes, giving an accurate impression of their relative average sizes. These archetypal outlines are expected to be of assistance to graders, who have hitherto operated by largely subjective criteria. The plant position classifier has the disadvantage that it

cannot operate on information regarding the physical texture or aroma of the leaf: it may, however, be expected to perform better if also supplied with images of the underside of a leaf, where the leaf's midrib is much more easily visible.

The historical context, economic framework, technical need, available image processing resources, data acquisition methodology and classification theory for the development of algorithms for these classifiers have all been covered in detail in this dissertation. The results achieved by the machine vision colour classifier challenge the accuracy of human graders, while the plant position classifier, given its limitations, performs about as well as a human expert working with visual information only, but not as well as a human grader with full access to the leaf.

Appendix A

Colour Classifier Statistics

A.1 Listing of all colour classifier measurements

Table A.1: **Feature values for colour classification, with class means**

Feature values for all data used in colour classification																
Data	Col	Tr/Tst	R md	$\bar{R}$	R var	R skew	R kurt	G md	$\bar{G}$	G var	G skew	G kurt	I md	$\bar{I}$	I var	I skew	I kurt	Rggdnss
2e1	plc	1	185	179.79	20.70	-0.97	5.73	124	128.65	19.45	-0.21	3.67	121	125.68	17.69	-0.11	3.85	0.0087
2e10	plc	2	189	181.77	17.69	-0.82	4.11	129	127.33	15.97	-0.13	3.92	125	123.52	14.17	0.04	4.60	0.0072
2e11	plc	1	187	184.53	18.44	-0.89	4.46	140	133.44	15.76	-0.19	3.42	132	128.39	14.34	-0.07	3.95	0.0055
2e12	plc	2	200	187.45	21.62	-1.19	4.37	137	134.95	19.07	-0.63	3.67	132	129.35	16.97	-0.67	4.06	0.0117
2e15	plc	1	191	181.20	17.80	-0.89	4.56	131	126.24	16.13	-0.25	3.66	127	123.37	14.64	-0.08	4.14	0.0095
2e16	plc	2	192	179.37	20.46	-1.21	5.00	136	129.56	15.32	-0.46	4.79	130	124.63	14.62	-0.26	5.12	0.0147
2e17	plc	1	203	184.12	23.05	-1.08	4.85	146	133.34	21.39	-0.52	3.50	134	128.02	19.01	-0.44	3.65	0.0096
2e2	plc	2	191	184.06	18.43	-1.37	7.08	136	132.84	16.56	-0.64	4.71	136	129.32	14.87	-0.57	5.02	0.0063
2e20	plc	1	187	177.81	21.56	-1.74	9.21	128	127.81	16.87	-0.90	6.91	126	124.06	15.38	-0.72	7.04	0.0241
2e3	plc	1	185	179.96	19.89	-0.71	3.76	129	125.33	17.68	-0.21	3.24	124	121.89	16.26	-0.08	3.49	0.0100
2e4	plc	1	192	181.20	19.27	-0.78	3.90	128	128.68	17.64	-0.20	3.19	130	124.06	15.74	-0.12	3.45	0.0063
2e5	plc	2	180	177.54	24.40	-2.55	15.63	131	128.69	18.14	-1.78	12.81	124	123.59	17.10	-1.65	13.15	0.0177
2e6	plc	1	189	181.31	21.07	-0.92	4.78	136	131.52	20.05	-0.28	3.41	132	126.75	17.96	-0.20	3.48	0.0073
2e61	plc	2	204	193.93	20.59	-1.84	9.71	140	145.27	19.81	-0.93	5.52	135	140.43	17.78	-0.85	5.39	0.0143
2e62	plc	1	191	182.08	21.22	-1.20	4.91	136	135.17	17.93	-0.55	3.88	136	130.06	15.84	-0.46	4.21	0.0080
2e63	plc	2	187	176.76	21.07	-1.46	6.52	132	128.97	18.13	-0.81	5.10	132	125.09	16.16	-0.67	5.54	0.0060
2e64	plc	1	195	183.02	19.94	-1.42	6.87	143	138.76	19.47	-0.59	4.25	142	133.71	17.28	-0.54	4.43	0.0044
2e65	plc	2	186	180.56	19.82	-1.22	6.32	133	128.49	18.11	-0.60	4.47	126	127.05	16.85	-0.49	4.41	0.0050
2e7	plc	1	187	178.63	24.01	-0.92	4.26	138	127.53	18.78	-0.41	3.78	129	123.70	17.14	-0.34	3.95	0.0221
2e8	plc	1	193	189.81	15.68	-1.26	5.86	139	136.47	14.55	-0.24	3.99	134	131.20	12.97	-0.02	4.96	0.0129
2e9	plc	1	191	191.01	15.74	-1.22	6.53	141	138.61	14.57	-0.34	4.09	134	133.37	12.86	-0.19	4.85	0.0087
3e1	plc	2	167	163.22	18.71	-0.29	2.95	116	110.14	18.44	0.18	3.37	116	109.89	16.56	0.34	3.98	0.0049
3e10	plc	1	189	174.51	19.26	-0.79	4.34	124	121.54	18.56	-0.08	3.27	128	119.66	16.60	0.02	3.70	0.0231
3e11	plc	2	183	167.74	22.85	-0.88	3.90	119	117.14	22.13	-0.36	3.50	120	115.72	19.53	-0.18	3.78	0.0140
3e12	plc	1	183	165.92	27.73	-0.93	4.39	123	113.66	24.14	-0.44	3.44	121	112.60	21.67	-0.38	3.66	0.0079
3e13	plc	2	186	175.13	20.03	-0.69	3.38	122	122.80	18.48	-0.20	3.01	125	120.00	16.58	-0.08	3.38	0.0101
3e4	plc	1	184	168.52	22.76	-1.37	6.06	119	116.73	19.67	-0.37	4.53	121	115.16	17.92	-0.15	5.26	0.0457
3e6	plc	2	186	174.55	21.90	-1.32	6.24	127	120.52	19.43	-0.46	4.17	124	118.64	17.21	-0.39	4.82	0.0211
3e7	plc	1	185	160.60	28.10	-0.73	3.01	117	110.93	26.38	-0.28	2.62	121	111.55	23.13	-0.25	2.81	0.0195
3e8	plc	1	173	160.90	25.18	-0.59	3.23	115	110.71	22.55	-0.17	3.08	114	110.06	20.35	-0.11	3.35	0.0207
111	lem	1	156	146.77	17.60	-0.80	4.00	102	100.45	17.54	0.05	3.58	99	96.59	16.24	0.24	4.43	0.0027
112	lem	2	160	150.44	18.47	-2.84	20.25	104	104.56	16.45	-0.85	8.85	103	99.38	15.25	-0.58	10.63	0.0021
113	lem	1	153	148.85	15.51	-0.85	4.88	102	99.60	15.08	0.12	5.18	98	95.14	14.08	0.52	6.94	0.0031
114	lem	2	159	156.93	16.72	-0.88	4.62	111	111.87	16.41	-0.28	3.89	107	105.90	14.77	-0.05	4.92	0.0023

Feature values for all data used in colour classification *(cont from prev page)*

Data	Col	Tr/Tst	R md	$\bar{R}$	R var	R skew	R kurt	G md	$\bar{G}$	G var	G skew	G kurt	I md	$\bar{I}$	I var	I skew	I kurt	Ragged
2i1	lem	1	169	162.56	20.76	-0.91	4.07	119	110.83	17.86	-0.56	4.12	118	110.05	16.19	-0.37	4.38	0.0060
2i10	lem	2	178	162.80	24.54	-0.76	3.61	116	108.19	21.27	-0.33	3.52	112	107.67	19.28	-0.22	3.73	0.0106
2i11	lem	1	163	159.33	17.90	-0.52	4.30	107	104.77	15.14	0.30	5.44	108	104.22	13.84	0.68	7.01	0.0023
2i12	lem	2	184	175.35	18.00	-0.96	5.46	119	117.75	15.80	-0.21	4.35	117	115.94	13.99	0.00	5.26	0.0043
2i13	lem	1	190	179.47	21.12	-1.00	4.75	128	122.26	18.86	-0.42	3.64	121	118.33	16.34	-0.31	4.06	0.0043
2i14	lem	1	192	183.56	20.42	-0.89	4.40	131	125.00	18.56	-0.14	3.72	126	122.25	16.33	-0.07	4.04	0.0030
2i15	lem	1	184	176.04	21.13	-0.94	4.19	123	122.27	18.52	-0.34	3.47	122	119.40	16.41	-0.20	3.88	0.0146
2i16	lem	2	183	167.13	21.10	-0.59	3.62	110	112.04	18.97	-0.17	3.69	114	110.47	16.95	0.03	4.20	0.0072
2i17	lem	1	158	167.89	23.20	-0.19	2.61	115	115.07	20.40	0.03	2.97	102	112.23	18.44	0.14	3.24	0.0253
2i18	lem	2	184	177.44	19.70	-0.79	4.52	119	119.19	17.64	-0.17	3.73	119	117.54	15.50	-0.06	4.34	0.0104
2i19	lem	1	179	167.49	20.95	-0.69	3.77	116	112.17	18.21	-0.21	3.65	114	109.56	16.31	-0.02	4.27	0.0149
2i2	lem	2	182	163.93	20.62	-0.93	4.18	116	110.79	16.78	-0.35	4.14	120	110.40	15.55	-0.20	4.44	0.0103
2i20	lem	1	180	163.55	22.30	-0.76	3.81	116	110.30	19.35	-0.34	3.53	118	108.84	17.47	-0.15	3.87	0.0060
2i3	lem	2	182	169.11	23.66	-0.87	4.51	121	112.61	21.10	-0.17	3.22	109	111.30	18.41	-0.07	3.86	0.0022
2i4	lem	1	180	161.41	21.76	-0.43	2.74	120	109.10	18.20	0.03	3.28	119	107.78	16.79	0.20	3.82	0.0071
2i5	lem	1	164	160.19	23.29	-0.65	3.64	118	107.51	20.22	-0.23	3.45	118	107.65	18.44	-0.14	3.55	0.0050
2i6	lem	1	168	159.93	24.00	-0.73	3.37	110	105.77	20.56	-0.38	3.39	110	106.61	18.62	-0.29	3.57	0.0048
2i7	lem	2	170	162.45	20.93	-0.52	3.42	111	107.42	17.25	-0.02	3.57	109	105.87	15.53	0.15	4.27	0.0021
2i8	lem	1	143	139.59	23.48	-0.46	3.49	84	88.97	19.54	0.19	3.90	91	90.11	17.73	0.33	4.70	0.0034
2i9	lem	2	179	164.95	22.64	-0.80	3.59	119	110.08	19.26	-0.26	3.54	118	109.47	17.41	-0.10	3.96	0.0165
3i1	lem	1	179	160.61	24.00	-0.92	4.26	119	110.77	20.44	-0.43	3.71	117	108.96	18.26	-0.24	4.11	0.0204
3i2	lem	2	137	138.77	24.06	-0.12	2.76	96	95.17	20.85	0.26	3.48	104	95.15	19.11	0.47	4.18	0.0235
3i3	lem	1	178	155.43	28.96	-0.48	2.86	122	102.60	26.22	-0.03	2.89	116	102.49	23.67	0.09	3.28	0.0275
3i4	lem	2	145	146.61	22.89	-0.43	3.22	104	102.05	20.14	0.19	4.13	103	101.39	18.61	0.52	4.98	0.0344
3i5	lem	1	156	153.74	19.19	-0.75	4.39	110	106.28	17.77	0.11	4.94	106	105.76	16.46	0.46	5.93	0.0386
3i6	lem	1	169	153.48	25.85	-0.77	3.57	109	104.24	22.26	-0.18	3.63	111	105.21	20.18	-0.02	4.09	0.0384
3i7	lem	1	150	148.50	20.57	-0.34	3.01	99	99.58	19.68	0.37	4.07	101	99.86	17.80	0.71	5.30	0.0256
3i8	lem	2	153	147.67	21.59	-0.71	3.93	97	98.24	17.94	0.12	5.16	102	98.97	16.59	0.39	6.04	0.0129
3i9	lem	1	150	145.19	23.14	-0.70	3.68	100	95.74	20.37	0.12	4.09	104	98.08	18.60	0.32	4.89	0.0302
3if1	lem	2	133	131.43	21.43	-0.33	3.40	83	83.40	19.03	0.94	6.32	85	87.50	17.89	1.20	7.43	0.0393
3if3	lem	1	147	135.15	30.70	-0.69	3.16	96	89.50	24.88	0.01	3.41	104	92.33	23.15	0.15	3.91	0.0347
3if4	lem	2	144	139.23	25.62	-0.84	3.89	95	92.82	22.37	0.36	4.77	101	96.20	21.01	0.61	5.42	0.0474
3if5	lem	1	150	137.17	31.18	-0.74	3.14	102	91.76	24.80	-0.12	3.18	103	94.36	22.98	0.00	3.65	0.0366
3ig1	lem	2	122	124.95	23.20	-0.01	2.74	77	82.88	20.86	0.46	3.69	87	83.25	18.58	0.76	4.85	0.0071
3ig2	lem	1	131	127.74	22.33	-0.42	3.76	98	90.47	19.12	0.35	5.17	87	89.08	17.61	0.79	6.62	0.0111
3ig3	lem	1	146	137.07	27.35	-0.40	2.56	111	94.00	21.47	-0.07	2.84	106	93.24	20.02	0.04	3.18	0.0064
3ig4	lem	1	146	146.84	24.48	-0.70	4.03	107	105.93	18.53	-0.11	4.28	99	102.51	17.23	0.22	5.07	0.0189
3ig5	lem	2	135	129.37	25.56	-0.19	2.78	103	91.28	21.78	0.26	4.16	91	90.12	20.11	0.70	5.23	0.0404
4i1	lem	1	132	113.85	28.54	-0.54	3.02	72	73.85	24.22	0.50	4.21	83	77.63	22.63	0.64	4.94	0.0385
4i10	lem	2	152	142.74	20.48	-0.52	3.31	101	96.79	19.29	0.08	4.11	99	99.47	17.82	0.33	4.84	0.0348
4i2	lem	1	119	110.74	29.59	-0.29	2.84	75	71.52	25.95	0.68	4.25	76	76.92	24.50	0.76	4.71	0.0540
4i3	lem	2	137	128.71	24.21	-0.56	3.61	90	89.69	22.09	0.32	4.35	99	92.09	20.62	0.53	5.12	0.0679
4i4	lem	1	144	126.75	31.83	-0.78	3.36	81	82.98	25.53	-0.10	3.50	97	86.82	23.89	-0.02	3.90	0.0355
4i5	lem	2	148	141.57	24.10	-0.10	2.38	116	98.32	23.59	0.21	2.87	113	99.96	21.53	0.40	3.41	0.0359
4i6	lem	1	139	140.80	20.37	-0.43	3.79	92	92.18	19.08	0.76	5.76	92	95.39	18.06	1.01	6.60	0.0442
4i7	lem	1	149	138.65	25.04	-0.35	2.84	96	93.04	21.74	0.16	3.57	100	95.82	20.14	0.34	4.09	0.0558
4i8	lem	1	128	121.35	21.25	-0.47	3.62	85	79.92	19.65	0.63	6.05	89	83.79	18.36	0.89	7.13	0.0315
4i9	lem	2	175	153.79	22.80	-0.40	2.69	122	108.16	21.96	0.03	3.23	125	109.13	20.30	0.26	3.63	0.0423
5i1	lem	1	135	134.89	22.38	-0.13	2.72	96	91.02	21.79	0.48	3.77	90	95.13	19.70	0.75	4.76	0.0584
5i10	lem	2	172	157.02	23.21	-0.52	3.01	118	110.17	22.14	-0.19	3.11	117	111.17	20.25	-0.01	3.42	0.0495
5i2	lem	1	149	144.97	17.66	-0.23	3.13	102	101.18	18.48	0.66	4.63	102	104.26	17.60	0.87	5.22	0.0220
5i4	lem	2	142	123.77	30.94	-0.44	2.57	92	83.83	26.29	0.04	2.93	94	87.23	24.06	0.16	3.39	0.0387
5i5	lem	1	136	119.33	33.15	-0.30	2.31	97	78.72	27.87	0.35	3.29	85	83.13	26.33	0.44	3.66	0.0460
5i6	lem	2	125	114.55	26.60	-0.17	2.76	84	78.74	24.94	0.67	4.36	89	83.76	23.56	0.83	4.94	0.1166
5i7	lem	1	140	116.88	28.66	-0.34	2.44	95	77.67	26.00	0.40	3.56	92	81.94	24.11	0.55	4.12	0.0312
5i8	lem	1	143	130.65	26.72	-0.68	3.35	92	87.35	23.61	0.07	3.91	94	90.80	21.92	0.27	4.52	0.0814
5i9	lem	1	144	118.42	32.54	-0.38	2.27	99	78.03	28.67	0.25	3.11	102	82.63	26.76	0.35	3.49	0.0369
1o3	ora	2	111	114.23	20.10	-0.27	3.47	63	68.30	19.34	0.85	6.59	69	69.51	18.50	1.15	7.69	0.0113
1o4	ora	1	114	105.66	19.23	-0.24	3.84	62	61.70	20.12	1.18	7.62	64	63.46	19.16	1.54	9.17	0.0111
2o10	ora	2	132	130.47	23.76	-0.21	2.79	70	75.50	20.42	0.63	4.50	75	81.25	18.39	0.77	5.46	0.0122
2o12	ora	1	133	140.32	28.70	-0.52	3.08	86	89.11	23.40	-0.12	3.16	92	90.99	21.24	0.01	3.58	0.0173
2o13	ora	2	142	136.19	22.64	-0.41	3.21	92	85.97	19.79	0.37	4.49	90	88.24	17.95	0.62	5.58	0.0130
2o14	ora	1	120	111.51	27.35	-0.38	2.76	70	62.61	20.89	0.76	5.45	76	68.96	19.52	0.81	6.15	0.0138
2o15	ora	2	160	141.97	31.00	-0.79	3.01	108	88.09	25.33	-0.34	3.02	104	90.68	22.65	-0.24	3.43	0.0052
2o16	ora	1	136	136.63	22.53	-0.55	4.22	83	84.24	19.32	0.70	5.69	87	87.76	17.84	0.90	6.88	0.0168
2o17	ora	1	140	125.86	28.50	-0.40	2.80	83	75.15	23.00	0.35	3.85	86	79.80	21.02	0.46	4.47	0.0155

Feature values for all data used in colour classification *(cont from prev page)*

Data	Col	Tr/Tst	R md	$\overline{R}$	R var	R skew	R kurt	G md	$\overline{G}$	G var	G skew	G kurt	I md	$\overline{I}$	I var	I skew	I kurt	Ragged
2o18	ora	1	135	123.52	24.70	-0.25	2.69	68	74.33	20.89	0.52	4.11	76	78.99	18.75	0.68	5.17	0.0101
2o19	ora	2	134	126.95	28.93	-0.52	3.02	81	77.83	24.48	0.35	3.94	84	82.02	22.41	0.49	4.66	0.0328
2o2	ora	1	120	122.50	30.17	0.03	2.49	63	71.23	27.03	0.76	3.81	70	76.61	24.52	0.90	4.50	0.0518
2o20	ora	2	118	114.78	21.83	-0.26	3.47	71	68.73	19.97	1.39	8.22	72	74.11	18.77	1.58	9.35	0.0397
2o4	ora	1	141	137.89	23.88	-0.83	4.13	87	85.57	20.11	0.24	4.44	91	89.03	18.35	0.35	5.25	0.0196
2o5	ora	2	149	139.42	23.06	-0.61	3.52	86	83.01	19.52	0.50	4.72	90	87.91	17.68	0.65	5.70	0.0135
2o6	ora	1	122	121.40	27.03	-0.18	2.68	69	69.15	22.39	0.70	4.58	74	75.66	20.69	0.76	5.16	0.0342
2o8	ora	2	124	120.18	23.97	-0.15	3.11	66	63.59	20.88	0.90	5.57	72	71.19	19.05	0.95	6.26	0.0087
3o1	ora	1	135	130.97	24.20	-0.56	3.19	79	80.34	20.11	0.35	4.32	82	84.85	18.44	0.49	5.18	0.0153
3o10	ora	1	137	129.47	24.46	-0.47	3.73	78	78.89	19.57	0.33	4.51	85	83.14	18.32	0.43	5.12	0.0268
3o2	ora	1	136	128.73	25.84	-0.51	3.15	85	78.02	20.90	0.37	4.41	88	83.26	19.38	0.45	5.10	0.0101
3o3	ora	2	137	126.61	21.31	-0.21	3.14	70	75.59	19.12	1.08	6.56	78	80.90	17.91	1.30	7.79	0.0352
3o4	ora	1	128	128.05	23.00	-0.42	3.35	77	77.52	19.06	0.57	5.24	83	82.47	17.57	0.71	6.22	0.0068
3o5	ora	2	140	132.50	21.47	-0.47	3.55	81	79.74	18.85	0.80	5.95	85	84.28	17.51	0.99	7.02	0.0257
3o6	ora	1	136	128.38	27.20	-0.23	2.51	81	76.75	22.10	0.41	3.61	84	81.99	20.27	0.49	4.14	0.0131
3o7	ora	2	148	135.80	23.52	-0.65	3.48	84	82.57	19.07	0.27	4.68	88	87.33	17.41	0.39	5.61	0.0135
3o8	ora	1	135	130.28	25.15	-0.24	2.81	82	80.06	21.26	0.57	4.32	86	84.63	19.60	0.72	5.11	0.0421
3o9	ora	2	130	129.88	23.07	-0.50	3.61	76	76.40	18.68	0.62	5.31	81	81.71	17.31	0.72	6.25	0.0119
3of1	ora	1	135	125.90	29.49	-0.39	2.77	80	80.12	24.15	0.23	3.15	83	83.98	22.13	0.31	3.63	0.0241
3of2	ora	1	142	131.20	28.23	-0.65	2.98	92	84.72	22.57	0.15	3.55	94	88.29	20.55	0.27	4.32	0.0366
3of3	ora	1	117	117.48	28.69	-0.34	2.91	72	73.88	23.53	0.53	4.19	75	78.06	21.89	0.64	4.96	0.0361
3of4	ora	2	148	131.92	26.70	-0.17	2.42	83	82.78	22.23	0.35	3.25	80	86.88	20.10	0.48	3.92	0.0152
3of5	ora	1	122	109.63	28.12	-0.51	3.13	68	64.64	22.90	0.72	5.29	73	69.91	21.37	0.83	6.13	0.0262
3og1	ora	2	103	99.90	22.06	0.21	3.36	52	56.14	19.15	1.94	10.28	64	63.37	18.05	1.99	10.99	0.0266
3og2	ora	1	98	100.35	25.64	0.51	3.19	53	58.56	23.66	1.85	7.72	56	65.53	22.46	1.88	8.15	0.0870
3og4	ora	2	118	110.93	20.50	0.06	2.99	60	63.72	18.73	1.11	6.51	71	69.61	16.81	1.37	8.41	0.0028
4o1	ora	1	106	103.35	26.29	0.00	3.13	52	60.03	23.44	1.63	7.60	67	66.85	22.20	1.67	8.10	0.0416
4o10	ora	2	118	102.22	30.54	-0.37	2.86	53	61.79	24.82	1.11	5.82	71	68.16	23.90	1.07	6.06	0.0581
4o2	ora	1	113	99.36	28.85	-0.62	3.22	63	57.86	21.92	1.06	7.30	70	64.52	21.25	0.94	7.28	0.0319
4o3	ora	1	76	83.21	29.20	0.53	3.21	38	47.98	26.42	2.04	8.12	48	55.67	25.39	1.93	7.95	0.0604
4o4	ora	1	117	95.60	32.10	-0.29	2.48	66	58.16	26.54	1.09	5.40	72	64.45	25.43	1.05	5.59	0.0671
4o5	ora	2	98	84.69	35.08	0.20	2.49	25	54.61	30.12	1.49	5.59	65	60.89	29.56	1.37	5.37	0.1123
4o7	ora	1	126	119.49	25.74	-0.38	3.21	78	76.57	24.69	1.12	5.72	83	82.03	23.44	1.24	6.22	0.0976
4o8	ora	2	119	97.96	34.20	-0.46	2.43	68	61.35	27.00	0.74	4.61	74	66.40	26.00	0.72	4.88	0.0695
4o9	ora	1	87	92.98	23.84	1.09	4.72	45	56.36	24.68	2.15	8.48	55	63.52	23.56	2.17	8.75	0.0995
5o1	ora	2	123	114.07	27.41	-0.35	2.60	75	70.96	23.35	0.58	4.36	83	75.93	21.59	0.73	5.12	0.0461
5o10	ora	1	108	93.99	33.11	-0.07	2.48	55	59.16	27.06	1.29	5.73	60	65.35	26.22	1.20	5.70	0.0533
5o2	ora	2	85	81.19	23.14	0.45	4.13	49	48.28	20.88	2.45	12.39	54	54.65	20.09	2.44	12.61	0.0180
5o3	ora	1	118	105.74	21.54	0.38	4.54	61	63.84	21.28	2.08	9.89	70	71.06	20.48	2.10	10.11	0.0694
5o4	ora	1	124	110.39	28.12	-0.17	2.37	77	66.87	24.75	0.68	4.40	83	73.43	22.92	0.78	5.02	0.0258
5o5	ora	1	77	87.01	27.24	0.37	3.09	40	51.57	22.91	2.08	9.52	50	58.74	22.23	1.95	9.18	0.0378
5o6	ora	2	101	86.53	31.73	-0.08	2.43	48	51.17	24.45	1.69	7.73	62	58.41	24.20	1.43	6.95	0.0622
5o7	ora	1	116	105.96	29.75	-0.51	3.01	72	63.89	23.99	1.01	6.00	73	69.51	22.91	1.03	6.44	0.0667
5o8	ora	2	97	96.65	26.34	0.05	3.19	55	57.76	23.92	1.78	7.91	61	65.55	23.00	1.68	7.81	0.0669
1r1	lma	1	57	64.45	23.72	1.14	5.55	2	21.98	26.76	2.39	10.47	17	31.59	23.41	2.87	13.81	0.0116
1r2	lma	2	59	54.72	23.77	1.21	6.81	2	17.11	26.29	2.89	13.16	15	26.32	22.75	3.37	17.43	0.0174
1r3	lma	1	66	63.22	23.80	0.57	4.70	2	23.06	23.52	2.59	13.10	16	31.31	21.50	2.88	15.58	0.0074
2r1	lma	2	72	70.77	22.52	0.46	7.10	32	38.93	20.43	3.22	16.77	46	46.57	20.11	2.81	15.37	0.0102
2r10	lma	1	68	72.63	23.39	1.30	5.87	29	40.48	22.49	2.89	13.54	40	48.16	21.72	2.77	13.11	0.0743
2r11	lma	1	78	88.40	23.84	0.61	3.79	38	47.83	21.14	2.48	12.09	55	55.69	20.23	2.45	12.23	0.0471
2r12	lma	1	65	70.17	22.77	1.08	6.03	31	38.68	21.18	3.23	15.89	41	46.81	20.68	2.99	14.76	0.0422
2r13	lma	2	74	74.51	21.85	1.41	7.18	37	41.39	21.69	3.17	15.61	47	49.12	21.00	3.05	15.12	0.0452
2r14	lma	1	81	83.61	24.96	0.87	4.55	40	47.01	23.69	2.62	11.39	48	55.29	22.85	2.52	11.12	0.0493
2r15	lma	2	75	84.24	23.44	0.85	4.40	38	49.42	21.88	2.50	11.46	46	57.23	21.22	2.38	11.02	0.0818
2r16	lma	1	87	82.47	18.82	1.22	7.32	44	44.93	18.29	3.53	19.82	53	53.34	17.53	3.40	19.34	0.0262
2r17	lma	2	72	71.54	18.85	1.97	10.55	33	39.89	18.81	3.97	21.92	43	47.78	18.25	3.83	21.19	0.0314
2r18	lma	1	69	79.61	25.70	1.04	4.53	28	43.95	24.82	2.30	9.50	38	52.59	23.93	2.20	9.22	0.0725
2r19	lma	2	80	81.75	20.12	0.91	5.69	38	43.64	18.57	3.14	17.56	51	51.54	17.79	3.07	17.49	0.0226
2r2	lma	1	86	83.02	20.77	0.94	6.02	41	45.67	20.33	2.87	14.69	50	53.80	19.42	2.81	14.71	0.0229
2r20	lma	1	90	93.51	20.69	0.47	4.13	47	52.20	18.70	2.23	11.91	55	60.03	17.87	2.18	11.97	0.0149
2r3	lma	1	60	64.49	21.44	2.09	10.00	29	35.14	22.05	3.64	18.02	37	43.18	21.32	3.50	17.35	0.0300
2r4	lma	2	100	100.01	21.36	-0.02	6.23	48	55.88	20.18	2.44	12.54	63	63.67	19.17	2.38	13.03	0.0261
2r5	lma	1	49	60.76	25.07	1.87	7.90	23	34.72	25.07	3.13	13.72	31	41.84	24.51	3.00	13.10	0.0570
2r6	lma	2	86	86.43	19.85	0.79	5.22	45	47.51	18.53	2.89	15.83	53	55.53	17.62	2.87	16.08	0.0203
2r7	lma	1	71	75.65	20.08	1.41	7.42	35	41.17	20.12	3.43	17.53	43	49.06	19.43	3.31	17.04	0.0490
2r8	lma	2	47	54.99	20.70	1.96	10.28	25	29.42	19.16	4.23	24.58	32	36.46	18.96	3.89	22.35	0.0223

Feature values for all data used in colour classification *(cont from prev page)*

Data	Col	Tr/Tst	R md	$\overline{R}$	R var	R skew	R kurt	G md	$\overline{G}$	G var	G skew	G kurt	I md	$\overline{I}$	I var	I skew	I kurt	Ragged
2r9	lma	1	62	67.46	19.52	2.08	10.76	27	36.18	19.32	4.13	23.25	38	44.03	18.71	3.97	22.32	0.0245
3r1	lma	2	68	73.63	24.24	1.06	4.89	32	40.93	21.58	2.91	14.23	38	48.81	21.06	2.74	13.47	0.0434
3r10	lma	1	61	70.47	26.49	0.71	4.06	25	33.77	21.88	2.69	13.55	37	42.38	21.58	2.40	12.21	0.0415
3r11	lma	1	74	76.97	24.48	0.94	4.87	33	41.69	23.17	2.72	12.37	41	49.84	22.46	2.57	11.80	0.0587
3r12	lma	1	73	73.27	24.35	0.53	3.75	34	39.16	20.02	2.68	14.01	44	46.94	19.65	2.41	12.82	0.0255
3r13	lma	2	78	86.92	21.39	0.96	5.49	40	46.39	21.45	2.81	13.57	50	54.78	20.34	2.80	13.79	0.0629
3r14	lma	1	67	74.10	22.54	1.16	5.72	28	39.49	20.73	3.19	16.21	39	48.36	20.37	2.93	14.80	0.0437
3r15	lma	2	63	75.10	23.50	0.98	4.90	23	37.30	21.69	2.78	13.65	40	45.68	20.97	2.64	13.13	0.0224
3r16	lma	1	76	73.55	24.24	1.04	5.69	34	39.18	22.34	2.97	14.10	42	47.72	21.87	2.72	13.01	0.0341
3r17	lma	2	86	85.29	22.56	0.65	4.11	37	45.58	20.36	2.58	12.73	48	54.23	19.62	2.45	12.26	0.0300
3r18	lma	1	69	76.25	24.13	0.99	4.70	31	43.20	22.73	2.56	11.61	41	51.18	22.05	2.45	11.20	0.0767
3r19	lma	2	72	83.73	27.11	0.70	3.58	32	45.93	24.87	2.25	9.41	41	54.56	24.07	2.12	9.04	0.1118
3r2	lma	1	72	77.43	21.69	0.95	5.64	31	40.10	20.08	3.24	16.90	44	49.49	19.48	3.00	15.71	0.0170
3r20	lma	1	66	74.68	25.64	0.77	3.88	25	36.59	21.82	2.57	12.78	35	45.72	21.39	2.29	11.38	0.0428
3r3	lma	1	91	90.14	24.55	0.54	3.24	36	47.00	20.68	2.10	10.35	48	55.76	19.60	2.06	10.54	0.0308
3r4	lma	2	75	76.24	29.70	0.30	3.37	31	37.03	23.25	2.28	11.00	41	45.65	23.10	1.90	9.54	0.0253
3r5	lma	1	88	83.08	28.06	0.26	2.98	28	41.71	23.18	1.84	8.79	44	50.03	22.50	1.68	8.46	0.0295
3r6	lma	2	73	76.49	21.61	1.02	5.66	27	37.42	20.61	3.09	16.21	41	46.88	19.82	2.93	15.50	0.0210
3r8	lma	1	83	70.99	26.83	0.50	3.66	28	36.44	20.81	2.49	12.82	43	45.02	20.88	2.05	10.81	0.0199
3r9	lma	2	72	75.16	21.53	1.36	6.81	31	38.47	20.55	3.42	17.64	42	47.46	19.87	3.23	16.70	0.0435
4r1	lma	1	44	66.35	27.21	0.88	3.96	21	40.14	23.44	2.47	11.09	32	46.60	23.17	2.30	10.41	0.0392
4r10	lma	2	76	76.56	25.48	0.93	4.35	32	46.20	24.39	2.41	10.12	46	53.58	23.81	2.28	9.64	0.0639
4r2	lma	1	56	65.95	26.76	1.23	4.89	28	41.22	24.69	2.38	9.96	40	47.77	24.18	2.28	9.65	0.0614
4r3	lma	1	80	71.74	25.15	0.73	4.38	41	41.33	22.58	2.87	13.29	43	48.92	22.18	2.63	12.28	0.0412
4r4	lma	1	91	89.26	21.47	0.48	4.22	49	50.57	19.96	2.38	11.93	58	58.38	19.15	2.36	12.12	0.0367
4r5	lma	2	89	83.71	26.01	0.47	3.59	41	48.33	23.36	2.10	9.60	51	55.46	22.60	2.06	9.65	0.0582
4r6	lma	1	87	92.41	21.40	0.90	5.21	48	53.16	21.57	2.57	11.90	54	61.08	20.65	2.55	11.98	0.0551
4r8	lma	2	73	79.95	21.49	1.46	7.00	39	46.99	22.66	2.85	12.78	47	54.86	21.98	2.73	12.31	0.0735
4r9	lma	1	85	82.50	23.43	0.89	4.66	41	49.40	22.76	2.45	10.94	55	56.23	22.02	2.43	10.97	0.0476
5r10	lma	2	44	52.85	24.27	1.99	8.25	22	32.19	23.48	3.25	14.76	30	38.80	23.28	3.07	13.81	0.0610
5r2	lma	1	45	54.74	24.79	2.05	8.47	25	34.72	24.82	3.03	13.03	33	41.46	24.55	2.84	12.14	0.0808
5r3	lma	2	47	60.48	26.36	1.89	7.29	23	38.44	26.86	2.75	10.89	32	45.29	26.47	2.63	10.36	0.1303
5r5	lma	1	50	61.11	24.46	1.44	6.07	25	35.77	21.64	3.17	15.49	31	43.14	21.59	2.87	13.80	0.0263
5r6	lma	1	48	57.16	25.22	1.83	7.56	28	37.39	25.18	2.88	12.26	34	43.71	24.84	2.74	11.63	0.0451
5r7	lma	1	39	55.01	25.68	1.59	6.43	20	32.99	22.68	3.10	14.67	28	39.34	22.59	2.88	13.51	0.0580
5r8	lma	2	70	73.08	24.25	1.37	6.13	36	44.79	24.67	2.64	11.06	50	52.54	24.09	2.50	10.52	0.0982
5r9	lma	1	62	65.24	22.37	1.85	8.07	27	39.38	22.54	3.05	13.84	40	46.55	22.26	2.90	13.09	0.0495
1s1	dma	2	32	35.70	26.46	1.91	8.81	2	9.22	35.31	2.38	8.06	11	17.30	24.89	3.74	18.29	0.0075
1s2	dma	1	2	24.18	26.54	2.43	10.90	2	6.88	40.89	1.87	5.45	2	12.29	24.93	3.99	19.95	0.0092
1s3	dma	2	28	27.98	25.20	2.42	11.34	2	7.13	38.70	2.01	6.20	10	13.69	23.64	4.12	21.60	0.0047
1s4	dma	1	27	31.58	24.35	2.11	10.51	2	6.98	33.76	2.63	9.42	10	14.47	21.81	4.47	25.62	0.0040
3s1	dma	2	32	45.14	21.82	2.51	11.76	21	29.22	19.83	3.91	21.46	25	34.96	20.04	3.56	18.97	0.0401
3s10	dma	1	31	42.00	23.12	2.59	11.80	18	27.20	20.87	3.90	20.50	25	33.09	21.20	3.52	17.89	0.0773
3s11	dma	1	33	41.53	21.40	2.70	13.40	17	25.44	19.79	4.21	23.78	24	31.32	20.05	3.80	20.77	0.0375
3s12	dma	1	37	41.51	19.02	3.00	16.04	20	24.84	18.04	4.71	28.65	27	30.85	18.10	4.37	26.00	0.0217
3s13	dma	2	43	51.66	23.18	2.31	10.19	22	30.75	22.61	3.50	16.62	27	37.53	22.45	3.27	15.36	0.0711
3s15	dma	1	22	38.78	24.14	2.55	11.47	16	26.14	21.91	3.78	19.36	20	31.40	22.33	3.41	16.86	0.0753
3s16	dma	2	29	41.84	23.24	2.65	11.92	18	28.13	22.17	3.69	18.37	24	33.63	22.27	3.41	16.60	0.0369
3s17	dma	1	38	44.31	20.73	2.69	13.57	21	27.38	18.98	4.29	24.85	25	33.58	19.31	3.83	21.36	0.0252
3s18	dma	2	32	45.13	25.59	2.52	10.31	19	30.05	24.38	3.41	15.46	25	35.75	24.42	3.22	14.39	0.0849
3s19	dma	1	27	38.28	19.79	2.71	14.61	18	24.50	17.19	4.76	30.89	22	30.21	17.71	4.10	25.30	0.0105
3s2	dma	2	36	44.97	21.21	2.67	13.29	20	29.01	19.57	4.19	23.69	27	35.07	19.82	3.76	20.59	0.0304
3s20	dma	1	33	44.01	22.96	2.47	10.91	21	28.97	21.66	3.56	17.62	26	34.82	21.80	3.27	15.82	0.0424
3s3	dma	1	34	48.78	28.75	2.05	7.69	20	33.87	28.19	2.78	10.75	25	39.15	28.07	2.63	10.12	0.1745
3s4	dma	1	35	42.13	22.39	3.06	14.56	21	28.18	22.23	3.92	19.59	27	33.94	22.15	3.68	18.12	0.0278
3s5	dma	2	29	41.11	22.62	2.80	13.08	20	27.48	20.82	4.00	21.20	24	32.92	21.14	3.66	18.85	0.0516
3s6	dma	1	32	41.71	22.04	2.68	12.53	18	26.81	20.67	3.95	20.84	24	32.52	20.87	3.64	18.76	0.0344
3s7	dma	2	23	35.54	19.50	2.87	15.64	19	23.14	16.99	4.94	32.45	22	28.46	17.52	4.29	26.85	0.0075
3s8	dma	1	30	38.30	19.14	3.45	19.29	20	25.25	18.25	4.78	29.50	25	31.22	18.49	4.27	25.36	0.0235
3s9	dma	2	31	42.65	22.02	2.69	12.58	20	28.44	20.66	3.91	20.63	24	33.88	20.84	3.60	18.49	0.0333
4s1	dma	1	30	39.82	23.26	2.68	12.22	18	26.32	22.26	3.72	18.32	23	31.63	22.46	3.45	16.66	0.0362
4s10	dma	2	41	52.62	24.58	2.46	9.71	27	37.78	25.08	3.02	12.18	32	43.53	24.83	2.92	11.74	0.0803
4s11	dma	1	44	56.00	24.46	2.00	8.67	27	33.94	24.01	3.35	15.29	35	40.70	23.74	3.15	14.29	0.0626
4s12	dma	1	26	46.45	28.65	2.09	7.60	18	33.27	27.57	2.77	10.72	23	38.41	27.56	2.64	10.14	0.0987
4s13	dma	1	30	42.95	28.79	2.41	8.97	20	32.32	28.50	2.85	10.92	24	37.02	28.53	2.73	10.38	0.1485
4s14	dma	2	23	37.14	22.11	3.07	15.05	17	25.23	21.79	4.07	20.91	21	29.92	21.83	3.85	19.59	0.0333

Feature values for all data used in colour classification *(cont from prev page)*

Data	Col	Tr/Tst	R md	$\overline{R}$	R var	R skew	R kurt	G md	$\overline{G}$	G var	G skew	G kurt	I md	$\overline{I}$	I var	I skew	I kurt	Ragged
4s15	dma	1	32	49.08	26.42	1.85	7.40	19	31.42	25.17	2.98	12.83	24	37.24	25.26	2.77	11.76	0.0747
4s16	dma	2	45	58.30	29.35	1.66	5.85	26	40.81	29.87	2.27	7.85	29	47.03	29.52	2.16	7.50	0.1455
4s17	dma	1	34	47.14	25.61	2.40	9.64	19	30.95	25.93	3.04	12.56	26	36.57	25.81	2.93	12.02	0.0704
4s18	dma	2	33	48.61	26.62	2.43	9.47	22	33.42	27.18	2.95	11.68	26	39.12	27.00	2.85	11.22	0.0897
4s20	dma	1	48	55.79	23.67	1.88	8.25	24	33.95	23.12	3.28	15.18	30	40.76	22.78	3.07	14.14	0.0511
4s3	dma	2	21	35.04	23.60	2.72	12.53	17	24.31	21.18	3.92	20.73	21	29.09	21.59	3.58	18.39	0.0224
4s4	dma	1	41	51.34	29.36	1.94	6.98	23	37.10	29.52	2.51	9.05	29	42.47	29.25	2.40	8.64	0.1524
4s5	dma	1	35	45.29	22.25	2.55	12.10	19	27.14	21.15	3.97	20.91	26	32.97	21.11	3.73	19.36	0.0786
4s6	dma	1	19	29.71	21.40	3.66	18.90	14	21.40	20.54	4.51	25.13	19	25.54	20.81	4.29	23.40	0.0427
4s7	dma	2	38	46.47	23.25	2.52	11.18	20	30.82	23.18	3.49	16.25	30	36.54	23.08	3.29	15.16	0.0472
4s9	dma	1	36	46.91	26.24	2.29	9.17	19	32.09	25.98	3.07	12.75	27	37.57	26.05	2.89	11.91	0.0491
5s1	dma	2	34	40.52	20.73	2.88	14.30	19	26.85	19.55	4.12	22.79	24	32.46	19.70	3.79	20.40	0.0260
5s10	dma	1	47	46.38	26.95	2.03	8.20	20	32.42	26.10	2.95	12.15	25	37.98	26.04	2.75	11.26	0.0974
5s2	dma	2	36	43.65	21.89	2.96	14.20	23	29.48	21.83	3.91	19.53	30	35.60	21.63	3.68	18.18	0.0360
5s4	dma	1	17	39.86	27.37	2.08	8.43	10	26.66	26.38	3.02	12.75	24	31.89	26.39	2.80	11.75	0.0801
5s5	dma	2	31	44.12	28.90	2.28	8.35	19	33.55	28.39	2.74	10.41	25	38.11	28.38	2.63	9.92	0.1088
5s6	dma	1	30	41.85	27.28	2.48	9.67	19	31.98	26.86	2.95	12.00	23	36.53	26.70	2.85	11.52	0.0933
5s7	dma	1	33	43.44	25.70	2.46	10.06	19	30.67	25.51	3.14	13.35	24	36.22	25.30	2.98	12.55	0.0543
5s8	dma	1	22	43.51	27.05	1.91	7.64	18	28.54	24.17	3.16	14.50	21	34.01	24.30	2.89	12.98	0.0702
5s9	dma	1	42	49.23	21.38	2.61	12.90	26	30.88	21.15	3.95	20.58	34	37.41	20.94	3.68	18.92	0.0391
mean	ple		188.03	178.23	20.97	-1.11	5.53	130.50	127.06	18.71	-0.43	4.23	127.70	123.68	16.84	-0.32	4.58	0.0129
mean	tem		156.07	146.98	23.48	-0.61	3.75	104.43	99.32	20.66	0.06	4.01	104.21	99.74	18.95	0.26	4.68	0.0260
mean	ora		122.36	115.81	26.23	-0.23	3.14	69.45	69.86	22.35	0.91	5.72	75.64	75.12	20.90	0.99	6.33	0.0352
mean	lma		70.29	74.41	23.52	1.08	5.76	31.29	40.62	22.07	2.84	13.79	41.71	48.56	21.36	2.71	13.37	0.0443
mean	dma		31.92	43.06	24.12	2.49	11.30	18.35	27.72	24.19	3.49	17.07	24.00	33.44	23.15	3.40	16.65	0.0575

A.2 Colour feature correlation matrices : within-groups, and over all data

Table A.2: **Pooled within-groups correlation matrix**

Matrix of colour feature correlations within groups

FEATURE	R md	$\overline{R}$	R std	R skew	R kurt	G md	$\overline{G}$	G std	G skew	G kurt	I md	$\overline{I}$	I std	I skew	I kurt	Rggdnss
R md	1.0000	0.8850	-0.2619	-0.5527	-0.0882	0.8458	0.8036	-0.3725	-0.4635	-0.1277	0.8922	0.8224	-0.4191	-0.5534	-0.2129	-0.2806
$\overline{R}$	0.8850	1.0000	-0.4088	-0.5110	-0.0863	0.8529	0.9450	-0.4666	-0.4896	-0.1725	0.8641	0.9484	-0.5293	-0.5555	-0.2523	-0.2463
R std	-0.2619	-0.4088	1.0000	-0.1436	-0.5147	-0.2896	-0.3865	0.7818	-0.3259	-0.4794	-0.2628	-0.3543	0.8784	-0.3800	-0.5298	0.4113
R skew	-0.5527	-0.5110	-0.1436	1.0000	0.5671	-0.3907	-0.4135	0.0168	0.7442	0.5102	-0.4529	-0.4286	0.0833	0.7275	0.4767	0.0714
R kurt	-0.0882	-0.0863	-0.5147	0.5671	1.0000	0.0170	-0.0264	-0.4338	0.5711	0.7700	-0.0125	-0.0648	-0.4739	0.5626	0.7786	-0.3446
G md	0.8458	0.8529	-0.2896	-0.3907	0.0170	1.0000	0.8921	-0.4094	-0.3897	-0.0720	0.9376	0.8915	-0.4189	-0.5106	-0.2109	-0.2105
$\overline{G}$	0.8036	0.9450	-0.3865	-0.4135	-0.0264	0.8921	1.0000	-0.4811	-0.4303	-0.1415	0.8738	0.9901	-0.4613	-0.5641	-0.3113	-0.1183
G std	-0.3725	-0.4666	0.7818	0.0168	-0.4338	-0.4094	-0.4811	1.0000	-0.3315	-0.6162	-0.4185	-0.4649	0.8642	-0.0947	-0.3514	0.3775
G skew	-0.4635	-0.4896	-0.3259	0.7442	0.5711	-0.3897	-0.4303	-0.3315	1.0000	0.8426	-0.4042	-0.4230	-0.1591	0.8472	0.6821	0.0085
G kurt	-0.1277	-0.1725	-0.4794	0.5102	0.7700	-0.0720	-0.1415	-0.6162	0.8426	1.0000	-0.0722	-0.1452	-0.4859	0.6379	0.7945	-0.2838
I md	0.8922	0.8641	-0.2628	-0.4529	-0.0125	0.9376	0.8738	-0.4185	-0.4042	-0.0722	1.0000	0.8923	-0.4085	-0.5541	-0.2336	-0.2067
$\overline{I}$	0.8224	0.9484	-0.3543	-0.4286	-0.0648	0.8915	0.9901	-0.4649	-0.4230	-0.1452	0.8923	1.0000	-0.4312	-0.5744	-0.3326	-0.0839
I std	-0.4191	-0.5293	0.8784	0.0833	-0.4739	-0.4189	-0.4613	0.8642	-0.1591	-0.4859	-0.4085	-0.4312	1.0000	-0.1781	-0.5246	0.6168
I skew	-0.5534	-0.5555	-0.3800	0.7275	0.5626	-0.5106	-0.5641	-0.0947	0.8472	0.6379	-0.5541	-0.5744	-0.1781	1.0000	0.8242	-0.1052
I kurt	-0.2129	-0.2523	-0.5298	0.4767	0.7786	-0.2109	-0.3113	-0.3514	0.6821	0.7945	-0.2336	-0.3326	-0.5246	0.8242	1.0000	-0.4552
Rggdnss	-0.2806	-0.2463	0.4113	0.0714	-0.3446	-0.2105	-0.1183	0.3775	0.0085	-0.2838	-0.2067	-0.0839	0.6168	-0.1052	-0.4552	1.0000

Table A.3: **Total correlation matrix for training data colour features**

Matrix of colour feature correlations over all training data															
FEATURE	R md	$\overline{R}$	R std	R skew	R kurt	G md	$\overline{G}$	G std	G skew	G kurt	I md	$\overline{I}$	I std	I skew	I kurt	Rggdnss
R md	1.0000	0.9907	-0.1717	-0.9411	-0.6459	0.9777	0.9726	-0.4574	-0.9402	-0.8101	0.9878	0.9804	-0.6162	-0.9511	-0.8378	-0.5342
$\overline{R}$	0.9907	1.0000	-0.2253	-0.9292	-0.6268	0.9802	0.9872	-0.4896	-0.9368	-0.8061	0.9868	0.9924	-0.6496	-0.9464	-0.8338	-0.5282
R std	-0.1717	-0.2253	1.0000	0.0068	-0.3096	-0.2056	-0.2502	0.7556	-0.0409	-0.2252	-0.1926	-0.2308	0.7858	-0.0517	-0.2302	0.4060
R skew	-0.9411	-0.9292	0.0068	1.0000	0.8115	-0.8867	-0.8826	0.3479	0.9484	0.8769	-0.9088	-0.8998	0.5089	0.9507	0.8846	0.4786
R kurt	-0.6459	-0.6268	-0.3096	0.8115	1.0000	-0.5473	-0.5410	0.0217	0.7268	0.8306	-0.5819	-0.5756	0.1128	0.7287	0.8350	0.1875
G md	0.9777	0.9802	-0.2056	-0.8867	-0.5473	1.0000	0.9901	-0.4631	-0.9241	-0.7775	0.9939	0.9895	-0.6162	-0.9361	-0.8122	-0.5030
$\overline{G}$	0.9726	0.9872	-0.2502	-0.8826	-0.5410	0.9901	1.0000	-0.4931	-0.9175	-0.7738	0.9879	0.9982	-0.6362	-0.9325	-0.8140	-0.4851
G std	-0.4574	-0.4896	0.7556	0.3479	0.0217	-0.4631	-0.4931	1.0000	0.2294	-0.0173	-0.4689	-0.4887	0.8823	0.3052	0.1256	0.5034
G skew	-0.9402	-0.9368	-0.0409	0.9484	0.7268	-0.9241	-0.9175	0.2294	1.0000	0.9401	-0.9301	-0.9237	0.4399	0.9828	0.9210	0.4413
G kurt	-0.8101	-0.8061	-0.2252	0.8769	0.8306	-0.7775	-0.7738	-0.0173	0.9401	1.0000	-0.7870	-0.7860	0.1993	0.9038	0.9454	0.2570
I md	0.9878	0.9868	-0.1926	-0.9088	-0.5819	0.9939	0.9879	-0.4689	-0.9301	-0.7870	1.0000	0.9912	-0.6169	-0.9455	-0.8250	-0.5098
$\overline{I}$	0.9804	0.9924	-0.2308	-0.8998	-0.5756	0.9895	0.9982	-0.4887	-0.9237	-0.7860	0.9912	1.0000	-0.6290	-0.9405	-0.8284	-0.4826
I std	-0.6162	-0.6496	0.7858	0.5089	0.1128	-0.6162	-0.6362	0.8823	0.4399	0.1993	-0.6169	-0.6290	1.0000	0.4398	0.2055	0.7191
I skew	-0.9511	-0.9464	-0.0517	0.9507	0.7287	-0.9361	-0.9325	0.3052	0.9828	0.9038	-0.9455	-0.9405	0.4398	1.0000	0.9444	0.4129
I kurt	-0.8378	-0.8338	-0.2302	0.8846	0.8350	-0.8122	-0.8140	0.1256	0.9210	0.9454	-0.8250	-0.8284	0.2055	0.9444	1.0000	0.1994
Rggdnss	-0.5342	-0.5282	0.4060	0.4786	0.1875	-0.5030	-0.4851	0.5034	0.4413	0.2570	-0.5098	-0.4826	0.7191	0.4129	0.1994	1.0000

A.3 Discriminant function analysis summary

Table A.4: **Summary of feature discriminant analysis**

Discriminatory potential of the colour features				
FEATURE	Wilks' Λ	Partial λ	F_{remove}	p-level
R md	0.00272	0.97203	0.94971	0.43754
$\overline{R}$	0.00310	0.85344	5.66704	0.00031
R std	0.00294	0.89943	3.68978	0.00700
R skew	0.00331	0.79933	8.28461	0.00001
R kurt	0.00363	0.72897	12.26933	0.00000
G md	0.00282	0.93779	2.18929	0.07361
$\overline{G}$	0.00326	0.81276	7.60245	0.00002
G std	0.00293	0.90456	3.48174	0.00973
G skew	0.00407	0.64986	17.78004	0.00000
G kurt	0.00324	0.81771	7.35654	0.00002
I md	0.00267	0.99332	0.22188	0.92581
$\overline{I}$	0.00278	0.95203	1.66272	0.16244
I std	0.00291	0.91058	3.24046	0.01426
I skew	0.00349	0.75945	10.45275	0.00000
I kurt	0.00303	0.87392	4.76096	0.00128
Rggdnss	0.00286	0.92512	2.67092	0.03492

Table A.5: **Summary of backwards stepwise discriminant analysis**

Order of removal of features, and discriminatory potentials					
Step number	Feature removed	Features left	Wilks' Λ	F-value	Partial λ
1	I md	15	0.00267	30.98261	0.99334
2	R md	14	0.00278	33.10596	0.96010
3	$\overline{I}$	13	0.00295	35.32255	0.93946
4	Rggdnss	12	0.00317	37.85069	0.93284
5	I std	11	0.00351	40.44268	0.90326
6	R std	10	0.00394	43.39521	0.88931
7	G std	9	0.00462	46.45352	0.85389
8	G md	8	0.00554	50.04818	0.83436
9	R skew	7	0.00671	54.74186	0.82500
10	I kurt	6	0.00897	59.21304	0.74767
11	I skew	5	0.01094	68.93661	0.81986
12	G kurt	4	0.01207	89.37536	0.90715
13	R kurt	3	0.01806	113.88425	0.66817
14	G skew	2	0.02639	188.20279	0.68436
15	$\overline{R}$	1	0.07890	429.04324	0.33443

Table A.6: **Summary of feature discriminant analysis (reduced subset)**

Discriminatory potential of the reduced feature subset				
FEATURE	Wilks' Λ	Partial λ	F_{remove}	p-level
$\overline{R}$	0.01021	0.65733	18.37567	0.00000
R kurt	0.00976	0.68723	16.04295	0.00000
$\overline{G}$	0.01201	0.55857	27.85748	0.00000
G skew	0.00981	0.68370	16.30768	0.00000
G kurt	0.00955	0.70268	14.91529	0.00000
I skew	0.00899	0.74593	12.00619	0.00000
I kurt	0.00897	0.74767	11.89635	0.00000

A.4 Squared Mahalanobis distances between classes

Table A.7: **Squared Mahalanobis distances between classes**

Squared Mahalanobis distances between class centroids					
CLASS	Pale Lemon	Lemon	Orange	Light Mahog	Dark Mahog
Pale Lemon	0.00	14.98	48.40	97.30	125.48
Lemon	14.98	0.00	15.23	53.11	87.85
Orange	48.40	15.23	0.00	17.91	54.00
Light Mahog	97.30	53.11	17.91	0.00	24.76
Dark Mahog	125.48	87.85	54.00	24.76	0.00

A.5 Colour classification functions

Table A.8: **Decision functions for three *a priori* distributions**

Decision functions for various *a priori* probabilities					
FEATURE	Pale Lemon	Lemon	Orange	Light Mahog	Dark Mahog
$\overline{R}$	1.3578	1.7197	2.3602	2.2551	1.6691
R kurt	1.9113	1.4127	2.7874	2.4553	5.6671
$\overline{G}$	0.3746	-0.4478	-1.5130	-1.4680	-0.9167
G skew	70.5226	33.8830	60.8630	83.9973	91.5746
G kurt	-10.2235	-4.7444	-7.8184	-10.3993	-12.3323
I skew	-32.2419	5.5353	-15.7416	-31.3705	-40.1302
I kurt	6.7815	0.8874	2.4362	5.0362	5.8264
Constant 1	-137.9386	-102.0708	-94.6376	-100.1496	-91.2179
Constant 2	-137.4145	-102.2673	-94.6572	-100.3187	-91.2046
Constant 3	-140.7669	-101.2965	-93.9665	-101.6467	-93.9380

A.6 Case-by-case presentation of test results

Table A.9: **Case-by-case unadjusted results for the colour classifier**

Colour classifier results based on training data class distribution																
DATA Obs	ple	lem	ora	lma	dma	ple	lem	ora	lma	dma	1	2	3	4	5
Class	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974
2e10 ple	3.13	13.15	43.85	88.42	127.05	0.986	0.014	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e12 ple	14.01	46.21	82.03	129.72	159.94	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e16 ple	6.40	17.96	62.76	114.83	144.74	0.994	0.006	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e2 ple	6.05	31.61	68.01	120.91	136.49	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e5 ple	171.74	209.27	261.57	337.44	311.77	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	dma	lma
2e61 ple	21.64	58.53	109.00	171.45	170.00	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	dma	lma
2e63 ple	11.58	32.30	77.08	132.49	148.90	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e65 ple	3.49	15.67	51.26	106.04	128.11	0.995	0.005	0.000	0.000	0.000	ple	lem	ora	lma	dma
*3e1 ple	12.60	5.41	21.03	57.90	100.61	0.013	0.987	0.000	0.000	0.000	lem	ple	ora	lma	dma
*3e11 ple	6.10	4.56	34.73	82.67	116.60	0.184	0.816	0.000	0.000	0.000	lem	ple	ora	lma	dma
3e13 ple	2.65	9.53	37.52	81.37	119.12	0.938	0.062	0.000	0.000	0.000	ple	lem	ora	lma	dma
3e6 ple	8.34	23.29	42.86	89.57	111.53	0.999	0.001	0.000	0.000	0.000	ple	lem	ora	lma	dma
*1l2 lem	233.34	256.38	258.13	316.67	249.47	0.999	0.000	0.000	0.000	0.001	ple	dma	lem	ora	lma
*1l4 lem	5.25	8.25	39.59	83.01	106.38	0.686	0.314	0.000	0.000	0.000	ple	lem	ora	lma	dma
2l10 lem	12.26	3.09	18.72	64.32	103.37	0.005	0.995	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l12 lem	8.53	9.57	27.23	73.12	107.13	0.450	0.549	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l16 lem	11.72	2.93	22.35	67.04	108.57	0.006	0.994	0.000	0.000	0.000	lem	ple	ora	lma	dma
*2l18 lem	8.26	12.63	27.67	71.20	109.54	0.812	0.188	0.000	0.000	0.000	ple	lem	ora	lma	dma
2l2 lem	9.17	3.12	24.52	71.34	106.99	0.023	0.977	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l3 lem	11.66	11.23	21.26	63.47	98.13	0.282	0.715	0.004	0.000	0.000	lem	ple	ora	lma	dma
2l7 lem	13.29	3.35	16.02	56.21	98.09	0.003	0.995	0.001	0.000	0.000	lem	ple	ora	lma	dma
2l9 lem	11.23	2.51	19.93	65.00	105.29	0.006	0.994	0.000	0.000	0.000	lem	ple	ora	lma	dma
3l2 lem	18.50	1.74	18.36	51.04	83.93	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
3l4 lem	24.73	7.67	34.93	72.82	107.46	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
3l8 lem	19.75	1.55	19.14	57.39	94.21	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
3lf1 lem	36.52	8.11	9.15	30.94	72.41	0.000	0.667	0.333	0.000	0.000	lem	ora	lma	ple	dma
3lf4 lem	24.91	2.40	14.93	48.56	81.26	0.000	0.998	0.002	0.000	0.000	lem	ora	ple	lma	dma
3lg1 lem	34.06	5.20	14.29	42.67	74.85	0.000	0.991	0.009	0.000	0.000	lem	ora	ple	lma	dma
3lg5 lem	53.59	24.63	55.76	92.15	123.18	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
4l10 lem	18.96	1.26	20.00	57.25	91.00	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4l3 lem	19.62	3.48	20.64	50.87	74.54	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4l5 lem	17.82	2.95	22.16	55.96	89.47	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
4l9 lem	16.97	6.12	34.92	75.09	111.39	0.002	0.998	0.000	0.000	0.000	lem	ple	ora	lma	dma
5l10 lem	10.95	3.80	32.53	75.93	110.36	0.013	0.987	0.000	0.000	0.000	lem	ple	ora	lma	dma
5l4 lem	26.21	5.08	15.28	47.66	73.82	0.000	0.995	0.005	0.000	0.000	lem	ora	ple	lma	dma
5l6 lem	30.51	6.88	15.48	37.15	61.96	0.000	0.989	0.011	0.000	0.000	lem	ora	ple	lma	dma
1o3 ora	59.17	15.80	9.96	32.05	70.71	0.000	0.060	0.940	0.000	0.000	ora	lem	lma	ple	dma
2o10 ora	59.38	24.30	4.89	28.62	74.89	0.000	0.000	1.000	0.000	0.000	ora	lem	lma	ple	dma
*2o13 ora	31.89	4.63	5.89	36.54	76.28	0.000	0.692	0.308	0.000	0.000	lem	ora	ple	lma	dma
2o15 ora	37.42	12.78	11.21	52.45	91.35	0.000	0.353	0.647	0.000	0.000	ora	lem	ple	lma	dma
2o19 ora	41.12	10.81	1.91	29.90	66.30	0.000	0.014	0.986	0.000	0.000	ora	lem	lma	ple	dma
2o20 ora	50.84	16.02	4.85	15.62	57.56	0.000	0.004	0.990	0.005	0.000	ora	lma	lem	ple	dma
2o5 ora	48.59	20.31	5.71	33.12	77.21	0.000	0.001	0.999	0.000	0.000	ora	lem	lma	ple	dma
2o8 ora	93.01	50.38	14.81	31.93	79.09	0.000	0.000	1.000	0.000	0.000	ora	lma	lem	dma	ple
3o3 ora	50.12	16.22	4.71	22.16	68.30	0.000	0.004	0.996	0.000	0.000	ora	lem	lma	ple	dma
3o5 ora	44.42	14.13	3.23	25.78	69.31	0.000	0.005	0.995	0.000	0.000	ora	lem	lma	ple	dma
3o7 ora	37.56	13.76	4.10	33.33	72.75	0.000	0.009	0.991	0.000	0.000	ora	lem	lma	ple	dma
3o9 ora	50.52	21.94	3.64	26.49	66.97	0.000	0.000	1.000	0.000	0.000	ora	lem	lma	ple	dma
3of4 ora	34.64	7.52	2.30	30.61	69.60	0.000	0.081	0.919	0.000	0.000	ora	lem	lma	ple	dma
*3og1 ora	70.58	32.09	8.61	6.64	49.90	0.000	0.000	0.243	0.757	0.000	lma	ora	lem	dma	ple
3og4 ora	56.57	19.66	2.07	14.15	55.48	0.000	0.000	0.997	0.003	0.000	ora	lma	lem	dma	ple
4o10 ora	53.99	22.27	2.58	12.34	42.29	0.000	0.000	0.991	0.009	0.000	ora	lma	lem	dma	ple
4o5 ora	62.05	30.70	11.99	12.54	31.84	0.000	0.000	0.531	0.469	0.000	ora	lma	lem	dma	ple
4o8 ora	51.04	19.83	7.17	21.91	46.80	0.000	0.002	0.997	0.001	0.000	ora	lem	lma	dma	ple
5o1 ora	42.60	9.35	2.51	24.82	57.99	0.000	0.038	0.962	0.000	0.000	ora	lem	lma	ple	dma
*5o2 ora	75.47	38.11	19.07	7.10	37.83	0.000	0.000	0.002	0.998	0.000	lma	ora	lem	dma	ple
*5o6 ora	74.43	41.55	13.19	9.74	39.11	0.000	0.000	0.133	0.867	0.000	lma	ora	dma	lem	ple
*5o8 ora	64.43	30.97	6.29	5.37	36.96	0.000	0.000	0.352	0.648	0.000	lma	ora	lem	dma	ple
1r2 lma	149.75	93.62	40.14	14.67	36.43	0.000	0.000	0.000	1.000	0.000	lma	dma	ora	lem	ple
2r1 lma	108.75	78.76	38.03	11.27	28.56	0.000	0.000	0.000	1.000	0.000	lma	dma	ora	lem	ple
2r13 lma	97.24	57.31	22.88	1.45	20.63	0.000	0.000	0.000	1.000	0.000	lma	dma	ora	lem	ple

Colour classifier results based on training data class distribution *(cont from prev page)*

DATA Obs		ple	lem	ora	lma	dma	ple	lem	ora	lma	dma	1	2	3	4	5
	Class	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974	p=.1184	p=.2434	p=.2040	p=.2368	p=.1974
2r15	lma	78.75	41.03	13.63	1.59	29.99	0.000	0.000	0.002	0.998	0.000	lma	ora	dma	lem	ple
2r17	lma	109.13	74.34	41.34	10.13	22.15	0.000	0.000	0.000	0.998	0.002	lma	dma	ora	lem	ple
2r19	lma	98.32	58.11	30.57	9.38	48.75	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
2r4	lma	83.34	53.81	18.78	8.72	38.71	0.000	0.000	0.006	0.994	0.000	lma	ora	dma	lem	ple
2r6	lma	88.74	50.51	24.98	7.61	46.96	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
2r8	lma	124.75	83.59	51.48	15.03	26.54	0.000	0.000	0.000	0.997	0.003	lma	dma	ora	lem	ple
3r1	lma	93.75	51.37	20.88	1.62	30.81	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r13	lma	96.34	53.20	18.54	3.36	37.71	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r15	lma	107.57	58.83	18.87	1.84	35.13	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r17	lma	92.71	49.17	15.51	2.30	40.92	0.000	0.000	0.001	0.999	0.000	lma	ora	dma	lem	ple
3r19	lma	91.47	47.11	11.76	3.04	34.57	0.000	0.000	0.011	0.989	0.000	lma	ora	dma	lem	ple
3r4	lma	116.59	70.95	23.53	10.81	45.61	0.000	0.000	0.001	0.999	0.000	lma	ora	dma	lem	ple
3r6	lma	111.71	63.44	23.99	3.49	39.95	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r9	lma	111.16	66.72	28.06	3.16	33.39	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
4r10	lma	80.50	42.30	14.72	2.66	22.55	0.000	0.000	0.002	0.998	0.000	lma	ora	dma	lem	ple
4r5	lma	75.97	35.30	9.20	2.57	32.16	0.000	0.000	0.030	0.970	0.000	lma	ora	dma	lem	ple
4r8	lma	91.34	54.74	21.01	5.15	18.49	0.000	0.000	0.000	0.999	0.001	lma	dma	ora	lem	ple
*5r10	lma	108.30	68.23	34.41	9.36	4.34	0.000	0.000	0.000	0.089	0.911	dma	lma	ora	lem	ple
*5r3	lma	101.69	64.25	31.47	13.92	7.37	0.000	0.000	0.000	0.043	0.957	dma	lma	ora	lem	ple
5r8	lma	87.24	51.27	20.28	5.29	13.89	0.000	0.000	0.000	0.988	0.011	lma	dma	ora	lem	ple
1s1	dma	162.80	109.91	67.87	44.91	33.41	0.000	0.000	0.000	0.004	0.996	dma	lma	ora	lem	ple
1s3	dma	193.96	149.07	120.12	99.68	65.12	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s1	dma	119.07	81.88	52.37	20.81	3.81	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s13	dma	120.81	81.72	43.04	15.66	2.95	0.000	0.000	0.000	0.002	0.998	dma	lma	ora	lem	ple
3s16	dma	128.42	92.97	58.89	28.80	0.65	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s18	dma	121.25	84.31	50.79	23.85	1.03	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s2	dma	127.87	91.94	60.20	26.13	4.67	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s5	dma	132.25	96.67	63.35	30.35	1.69	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s7	dma	161.71	117.92	95.89	56.45	37.26	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s9	dma	126.89	91.85	59.53	27.36	1.24	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s10	dma	113.22	82.04	53.13	32.01	5.41	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s14	dma	155.60	123.68	85.56	50.68	7.21	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s16	dma	92.23	58.01	33.61	21.54	11.05	0.000	0.000	0.000	0.006	0.994	dma	lma	ora	lem	ple
4s18	dma	121.64	85.70	52.55	31.94	5.63	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s3	dma	133.76	96.52	65.31	32.25	2.27	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s7	dma	124.61	90.20	55.09	27.29	1.39	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s1	dma	137.91	103.06	69.65	35.87	3.50	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s2	dma	146.47	116.50	78.24	45.68	6.77	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s5	dma	113.92	79.98	54.23	35.05	7.03	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple

Table A.10: **Case-by-case adjusted results for the colour classifier**

Colour classifier results based on "real world" class distribution

DATA Obs		ple	lem	ora	lma	dma	ple	lem	ora	lma	dma	1	2	3	4	5
	Class	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130
2e10	ple	3.13	13.15	43.85	88.42	127.05	0.664	0.336	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e12	ple	14.01	46.21	82.03	129.72	159.94	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e16	ple	6.40	17.96	62.76	114.83	144.74	0.811	0.189	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e2	ple	6.05	31.61	68.01	120.91	136.49	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e5	ple	171.74	209.27	261.57	337.44	311.77	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	dma	lma
2e61	ple	21.64	58.53	109.00	171.45	170.00	1.000	0.000	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e63	ple	11.58	32.30	77.08	132.49	148.90	0.998	0.002	0.000	0.000	0.000	ple	lem	ora	lma	dma
2e65	ple	3.49	15.67	51.26	106.04	128.11	0.854	0.146	0.000	0.000	0.000	ple	lem	ora	lma	dma
*3e1	ple	12.60	5.41	21.03	57.90	100.61	0.000	0.999	0.000	0.000	0.000	lem	ple	ora	lma	dma
*3e11	ple	6.10	4.56	34.73	82.67	116.60	0.006	0.994	0.000	0.000	0.000	lem	ple	ora	lma	dma
*3e13	ple	2.65	9.53	37.52	81.37	119.12	0.292	0.708	0.000	0.000	0.000	lem	ple	ora	lma	dma
3e6	ple	8.34	23.29	42.86	89.57	111.53	0.959	0.041	0.000	0.000	0.000	ple	lem	ora	lma	dma
*1l2	lem	233.34	256.38	258.13	316.67	249.47	0.998	0.001	0.000	0.000	0.001	ple	lem	dma	ora	lma
1l4	lem	5.25	8.25	39.59	83.01	106.38	0.056	0.944	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l10	lem	12.26	3.09	18.72	64.32	103.37	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
2l12	lem	8.53	9.57	27.23	73.12	107.13	0.022	0.978	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l16	lem	11.72	2.93	22.35	67.04	108.57	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
2l18	lem	8.26	12.63	27.67	71.20	109.54	0.105	0.895	0.000	0.000	0.000	lem	ple	ora	lma	dma

Colour classifier results based on "real world" class distribution *(cont from prev page)*

DATA Obs		ple	lem	ora	lma	dma	ple	lem	ora	lma	dma	1	2	3	4	5
	Class	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130
2i2	lem	9.17	3.12	24.52	71.34	106.99	0.001	0.999	0.000	0.000	0.000	lem	ple	ora	lma	dma
2i3	lem	11.66	11.23	21.26	63.47	98.13	0.011	0.984	0.005	0.000	0.000	lem	ple	ora	lma	dma
2i7	lem	13.29	3.35	16.02	56.21	98.09	0.000	0.999	0.001	0.000	0.000	lem	ora	ple	lma	dma
2i9	lem	11.23	2.51	19.93	65.00	105.29	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
3i2	lem	18.50	1.74	18.36	51.04	83.93	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
3i4	lem	24.73	7.67	34.93	72.82	107.46	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
3i8	lem	19.75	1.55	19.14	57.39	94.21	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
3if1	lem	36.52	8.11	9.15	30.94	72.41	0.000	0.690	0.310	0.000	0.000	lem	ora	lma	ple	dma
3if4	lem	24.91	2.40	14.93	48.56	81.26	0.000	0.999	0.001	0.000	0.000	lem	ora	ple	lma	dma
3ig1	lem	34.06	5.20	14.29	42.67	74.85	0.000	0.992	0.008	0.000	0.000	lem	ora	ple	lma	dma
3ig5	lem	53.59	24.63	55.76	92.15	123.18	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4i10	lem	18.96	1.26	20.00	57.25	91.00	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4i3	lem	19.62	3.48	20.64	50.87	74.54	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4i5	lem	17.82	2.95	22.16	55.96	89.47	0.000	1.000	0.000	0.000	0.000	lem	ora	ple	lma	dma
4i9	lem	16.97	6.12	34.92	75.09	111.39	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
5i10	lem	10.95	3.80	32.53	75.93	110.36	0.000	1.000	0.000	0.000	0.000	lem	ple	ora	lma	dma
5i4	lem	26.21	5.08	15.28	47.66	73.82	0.000	0.995	0.005	0.000	0.000	lem	ora	ple	lma	dma
5i6	lem	30.51	6.88	15.48	37.15	61.96	0.000	0.990	0.010	0.000	0.000	lem	ora	ple	lma	dma
1o3	ora	59.17	15.80	9.96	32.05	70.71	0.000	0.067	0.933	0.000	0.000	ora	lem	lma	ple	dma
2o10	ora	59.38	24.30	4.89	28.62	74.89	0.000	0.000	1.000	0.000	0.000	ora	lem	lma	ple	dma
*2o13	ora	31.89	4.63	5.89	36.54	76.28	0.000	0.714	0.286	0.000	0.000	lem	ora	ple	lma	dma
2o15	ora	37.42	12.78	11.21	52.45	91.35	0.000	0.376	0.624	0.000	0.000	ora	lem	ple	lma	dma
2o19	ora	41.12	10.81	1.91	29.90	66.30	0.000	0.015	0.985	0.000	0.000	ora	lem	lma	ple	dma
2o20	ora	50.84	16.02	4.85	15.62	57.56	0.000	0.005	0.994	0.001	0.000	ora	lem	lma	ple	dma
2o5	ora	48.59	20.31	5.71	33.12	77.21	0.000	0.001	0.999	0.000	0.000	ora	lem	lma	ple	dma
2o8	ora	93.01	50.38	14.81	31.93	79.09	0.000	0.000	1.000	0.000	0.000	ora	lma	lem	dma	ple
3o3	ora	50.12	16.22	4.71	22.16	68.30	0.000	0.004	0.996	0.000	0.000	ora	lem	lma	ple	dma
3o5	ora	44.42	14.13	3.23	25.78	69.31	0.000	0.006	0.994	0.000	0.000	ora	lem	lma	ple	dma
3o7	ora	37.56	13.76	4.10	33.33	72.75	0.000	0.010	0.990	0.000	0.000	ora	lem	lma	ple	dma
3o9	ora	50.52	21.94	3.64	26.49	66.97	0.000	0.000	1.000	0.000	0.000	ora	lem	lma	ple	dma
3of4	ora	34.64	7.52	2.30	30.61	69.60	0.000	0.089	0.911	0.000	0.000	ora	lem	lma	ple	dma
3og1	ora	70.58	32.09	8.61	6.64	49.90	0.000	0.000	0.737	0.263	0.000	ora	lma	lem	dma	ple
3og4	ora	56.57	19.66	2.07	14.15	55.48	0.000	0.000	0.999	0.000	0.000	ora	lma	lem	dma	ple
4o10	ora	53.99	22.27	2.58	12.34	42.29	0.000	0.000	0.999	0.001	0.000	ora	lma	lem	dma	ple
4o5	ora	62.05	30.70	11.99	12.54	31.84	0.000	0.000	0.908	0.092	0.000	ora	lma	lem	dma	ple
4o8	ora	51.04	19.83	7.17	21.91	46.80	0.000	0.002	0.998	0.000	0.000	ora	lem	lma	dma	ple
5o1	ora	42.60	9.35	2.51	24.82	57.99	0.000	0.042	0.958	0.000	0.000	ora	lem	lma	ple	dma
*5o2	ora	75.47	38.11	19.07	7.10	37.83	0.000	0.000	0.019	0.981	0.000	lma	ora	lem	dma	ple
5o6	ora	74.43	41.55	13.19	9.74	39.11	0.000	0.000	0.573	0.427	0.000	ora	lma	lem	dma	ple
5o8	ora	64.43	30.97	6.29	5.37	36.96	0.000	0.000	0.826	0.174	0.000	ora	lma	lem	dma	ple
1r2	lma	149.75	93.62	40.14	14.67	36.43	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
2r1	lma	108.75	78.76	38.03	11.27	28.56	0.000	0.000	0.000	1.000	0.000	lma	dma	ora	lem	ple
2r13	lma	97.24	57.31	22.88	1.45	20.63	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
2r15	lma	78.75	41.03	13.63	1.59	29.99	0.000	0.000	0.018	0.982	0.000	lma	ora	dma	lem	ple
2r17	lma	109.13	74.34	41.34	10.13	22.15	0.000	0.000	0.000	0.999	0.001	lma	dma	ora	lem	ple
2r19	lma	98.32	58.11	30.57	9.38	48.75	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
2r4	lma	83.34	53.81	18.78	8.72	38.71	0.000	0.000	0.047	0.953	0.000	lma	ora	dma	lem	ple
2r6	lma	88.74	50.51	24.98	7.61	46.96	0.000	0.000	0.001	0.999	0.000	lma	ora	lem	dma	ple
2r8	lma	124.75	83.59	51.48	15.03	26.54	0.000	0.000	0.000	0.999	0.001	lma	dma	ora	lem	ple
3r1	lma	93.75	51.37	20.88	1.62	30.81	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r13	lma	96.34	53.20	18.54	3.36	37.71	0.000	0.000	0.004	0.996	0.000	lma	ora	dma	lem	ple
3r15	lma	107.57	58.83	18.87	1.84	35.13	0.000	0.000	0.002	0.998	0.000	lma	ora	dma	lem	ple
3r17	lma	92.71	49.17	15.51	2.30	40.92	0.000	0.000	0.010	0.990	0.000	lma	ora	dma	lem	ple
3r19	lma	91.47	47.11	11.76	3.04	34.57	0.000	0.000	0.088	0.912	0.000	lma	ora	dma	lem	ple
3r4	lma	116.59	70.95	23.53	10.81	45.61	0.000	0.000	0.013	0.987	0.000	lma	ora	dma	lem	ple
3r6	lma	111.71	63.44	23.99	3.49	39.95	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
3r9	lma	111.16	66.72	28.06	3.16	33.39	0.000	0.000	0.000	1.000	0.000	lma	ora	dma	lem	ple
4r10	lma	80.50	42.30	14.72	2.66	22.55	0.000	0.000	0.018	0.982	0.000	lma	ora	dma	lem	ple
4r5	lma	75.97	35.30	9.20	2.57	32.16	0.000	0.000	0.215	0.785	0.000	lma	ora	lem	dma	ple
4r8	lma	91.34	54.74	21.01	5.15	18.49	0.000	0.000	0.003	0.997	0.000	lma	ora	dma	lem	ple
*5r10	lma	108.30	68.23	34.41	9.36	4.34	0.000	0.000	0.000	0.249	0.751	dma	lma	ora	lem	ple
*5r3	lma	101.69	64.25	31.47	13.92	7.37	0.000	0.000	0.000	0.133	0.866	dma	lma	ora	lem	ple
5r8	lma	87.24	51.27	20.28	5.29	13.89	0.000	0.000	0.004	0.993	0.003	lma	ora	dma	lem	ple
1s1	dma	162.80	109.91	67.87	44.91	33.41	0.000	0.000	0.000	0.013	0.987	dma	lma	ora	lem	ple
1s3	dma	193.96	149.07	120.12	99.68	65.12	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s1	dma	119.07	81.88	52.37	20.81	3.81	0.000	0.000	0.000	0.001	0.999	dma	lma	ora	lem	ple

Colour classifier results based on "real world" class distribution *(cont from prev page)*

DATA Obs	ple	lem	ora	lma	dma	ple	lem	ora	lma	dma	1	2	3	4	5
Class	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130	p=.0070	p=.5280	p=.3990	p=.0530	p=.0130
3s13 dma	120.81	81.72	43.04	15.66	2.95	0.000	0.000	0.000	0.007	0.993	dma	lma	ora	lem	ple
3s16 dma	128.42	92.97	58.89	28.80	0.65	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s18 dma	121.25	84.31	50.79	23.85	1.03	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s2 dma	127.87	91.94	60.20	26.13	4.67	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s5 dma	132.25	96.67	63.35	30.35	1.69	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s7 dma	161.71	117.92	95.89	56.45	37.26	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
3s9 dma	126.89	91.85	59.53	27.36	1.24	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s10 dma	113.22	82.04	53.13	32.01	5.41	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s14 dma	155.60	123.68	85.56	50.68	7.21	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s16 dma	92.23	58.01	33.61	21.54	11.05	0.000	0.000	0.000	0.021	0.979	dma	lma	ora	lem	ple
4s18 dma	121.64	85.70	52.55	31.94	5.63	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s3 dma	133.76	96.52	65.31	32.25	2.27	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
4s7 dma	124.61	90.20	55.09	27.29	1.39	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s1 dma	137.91	103.06	69.65	35.87	3.50	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s2 dma	146.47	116.50	78.24	45.68	6.77	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple
5s5 dma	113.92	79.98	54.23	35.05	7.03	0.000	0.000	0.000	0.000	1.000	dma	lma	ora	lem	ple

Appendix B

Plant Position Classifier Statistics

B.1 Discriminant function analysis summary

Table B.1: **Summary of forward stepwise discriminant analysis**

Order of addition of features, and discriminatory potentials				
FEATURE	Step no.	F-to-include	p-level	Wilks' Λ
width variance	1	81.50089	0.000000	0.333603
area	2	25.52438	0.000000	0.204831
error in fit to mean smoker template	3	13.63237	0.000000	0.153152
tip angle	4	5.26589	0.000144	0.135414
absolute length	5	6.28707	0.000019	0.117021
error in fit to mean tip template	6	3.26184	0.007463	0.108157
absolute width	7	2.82360	0.017349	0.100958
error in fit to mean leaf template	8	2.66751	0.023352	0.094556
error in fit to mean priming template	9	1.34814	0.245725	0.091413
error in fit to mean cutter template	10	1.45871	0.205122	0.088117

Table B.2: **Individual discriminatory value of features in the plant position classifier**

Discriminatory potential of individual features				
FEATURE	Wilks' Λ	Partial λ	F-to-remove	p-level
width variance	0.091407	0.096400	1.456313	0.205936
area	0.103369	0.085245	6.750421	0.000008
error in fit to mean smoker template	0.091989	0.957907	1.713751	0.133121
tip angle	0.092946	0.948045	2.137280	0.062644
absolute length	0.100359	0.878014	5.418417	0.000109
error in fit to mean tip template	0.093039	0.947095	2.178532	0.058102
absolute width	0.095730	0.920473	3.369536	0.006083
error in fit to mean leaf template	0.091473	0.963305	1.485617	0.196164
error in fit to mean priming template	0.091778	0.960107	1.620456	0.156282
error in fit to mean cutter template	0.091413	0.963946	1.458707	0.205122

B.2 Plant position classification functions

Table B.3: **Decision functions for the plant position classifier**

Decision functions for the six plant position groups						
FEATURE	primings	lugs	cutters	leaf	smokers	tips
width variance	6.20	6.22	6.24	6.18	6.17	6.14
area	-1894.59	-1866.81	-1832.00	-1815.30	-1923.60	-1808.28
error in fit to mean smoker template	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08
tip angle	1.83	1.75	1.81	1.78	1.85	1.78
absolute length	0.09	0.08	0.08	0.08	0.09	0.08
error in fit to mean tip template	0.27	0.27	0.27	0.27	0.27	0.27
absolute width	0.08	0.08	0.08	0.08	0.09	0.09
error in fit to mean leaf template	-0.66	-0.66	-0.66	-0.66	-0.66	-0.65
error in fit to mean priming template	0.68	0.68	0.68	0.68	0.68	0.67
error in fit to mean cutter template	-0.23	-0.23	-0.23	-0.23	-0.23	-0.23
constant	-2384.79	-2405.53	-2399.26	-2382.41	-2397.61	-2322.36

B.3 Case-by-case presentation of test results

Table B.4: **Squared Mahalanobis distances and *a posteriori* probabilities**

Plant position classification results													
Data	PlPos	priming	lug	cutter	leaf	smoker	tip	priming	lug	cutter	leaf	smoker	tip
*1	priming	13.6052	14.5014	10.2476	10.0642	10.5376	21.3080	0.057049	0.036445	0.305732	0.335091	0.264470	0.001212
2	priming	10.1442	11.6611	13.2572	13.9062	12.5590	21.1246	0.468431	0.219409	0.098776	0.071405	0.140045	0.001933
3	priming	7.0569	9.8493	10.2837	20.7881	21.2750	35.7063	0.690315	0.170884	0.137516	0.000720	0.000564	0.000000
4	priming	5.6447	9.2927	10.3433	21.4702	20.7632	33.6301	0.795097	0.128314	0.075883	0.000291	0.000414	0.000001
5	priming	10.5623	11.6373	17.4990	23.6689	28.6926	37.2264	0.618458	0.361311	0.019277	0.000882	0.000072	0.000001
6	priming	2.4391	3.1568	6.2914	12.7526	16.6070	31.1460	0.540315	0.377393	0.078727	0.003112	0.000453	0.000000
*7	priming	8.9748	7.8775	9.0146	9.5598	13.5227	30.8286	0.219274	0.379541	0.214952	0.163665	0.022564	0.000004
*8	priming	4.0763	3.9653	7.8728	12.6290	13.1917	32.8934	0.448171	0.473753	0.067150	0.006227	0.004700	0.000000
9	priming	6.8330	7.2007	8.7916	13.1746	16.1609	26.6803	0.442656	0.368320	0.166250	0.018578	0.004174	0.000022
10	priming	0.8959	1.7615	2.9468	7.5013	9.9817	22.3408	0.486669	0.315704	0.174535	0.017902	0.005179	0.000011
11	priming	5.9900	6.1417	9.7605	17.9472	21.7931	34.9317	0.480382	0.445303	0.072920	0.001217	0.000178	0.000000
12	priming	15.3070	19.7161	19.9802	31.5885	31.5819	42.5752	0.828132	0.091341	0.080043	0.000241	0.000242	0.000001
13	priming	4.1921	6.2975	9.0769	15.2518	18.9611	27.9359	0.694180	0.242269	0.060362	0.002754	0.000431	0.000005
*14	priming	7.9261	5.5247	7.5253	6.5139	10.4735	34.5658	0.127387	0.423228	0.155650	0.258093	0.035642	0.000000
15	priming	11.0265	11.6026	13.2409	16.7078	19.3460	21.9888	0.463305	0.347362	0.153119	0.027052	0.007233	0.001929
16	priming	2.8943	4.4634	9.8635	15.4605	18.2874	33.2335	0.671450	0.306401	0.020590	0.001254	0.000305	0.000600
17	priming	15.7986	18.3756	24.9648	24.4498	26.3588	37.3628	0.766732	0.211370	0.007838	0.010140	0.003904	0.000016
18	priming	4.9005	5.6531	8.4986	14.5825	17.3895	28.0882	0.537141	0.368695	0.088873	0.004243	0.001043	0.000005
19	priming	4.8598	7.3655	8.2771	15.0219	17.3112	27.2934	0.677964	0.193686	0.122788	0.004212	0.001341	0.000009
20	priming	9.4376	9.6956	17.4495	19.3430	25.7752	32.4278	0.525069	0.461507	0.009560	0.003709	0.000149	0.000005
21	priming	7.0250	10.1890	16.0777	20.1107	23.8859	34.8759	0.820988	0.168766	0.008883	0.001183	0.000179	0.000001
*22	priming	2.2365	2.2163	5.2698	9.8091	11.7726	28.9917	0.442324	0.446814	0.097071	0.010032	0.003759	0.000001
23	priming	2.1272	5.6009	8.9281	16.1553	14.8333	29.6107	0.825033	0.145266	0.027521	0.000742	0.001437	0.000001
24	priming	5.9451	9.8958	12.1147	18.5952	14.6141	34.2461	0.833781	0.115657	0.038137	0.001493	0.010930	0.000001
25	priming	10.5375	12.9903	17.8035	24.0176	25.8011	35.7488	0.756746	0.221984	0.020005	0.000895	0.000367	0.000003
*26	priming	11.6100	10.1353	11.4064	16.3153	21.6739	38.3933	0.232600	0.486226	0.257531	0.022124	0.001518	0.000000
27	priming	4.7350	5.6781	10.1711	18.2303	19.7179	33.9782	0.591095	0.368867	0.039015	0.000694	0.000330	0.000000
28	priming	4.1426	5.9565	6.7961	14.7841	9.3895	25.7983	0.572563	0.231168	0.151919	0.002799	0.041540	0.000011
29	priming	7.4061	12.4726	13.0217	21.8808	20.3955	32.0181	0.875679	0.069529	0.052835	0.000630	0.001323	0.000004
*30	priming	7.6627	7.9059	7.2734	10.0642	9.8850	14.9066	0.266154	0.235679	0.323339	0.080102	0.087612	0.007115
31	priming	3.1456	3.8150	4.0176	6.8313	8.3762	23.0785	0.385539	0.275884	0.249303	0.061056	0.028201	0.000018
*32	priming	5.2575	2.9164	2.8968	4.9377	11.5981	25.4803	0.115011	0.370778	0.374423	0.134954	0.004829	0.000005
*33	priming	6.3117	5.1363	5.5014	5.6517	8.9607	24.8629	0.167892	0.302168	0.251753	0.233524	0.044647	0.000016
34	priming	4.3240	5.0578	5.7623	9.7214	10.3761	29.4253	0.435566	0.301795	0.212191	0.029311	0.021129	0.000002
35	priming	5.7912	7.3526	11.7541	15.9470	17.2586	32.0212	0.658647	0.301719	0.033406	0.004105	0.002131	0.000001
36	lug	5.9109	3.9298	5.3407	8.4787	14.4449	32.1964	0.188196	0.506763	0.250280	0.052122	0.002639	0.000000
37	lug	8.3084	6.7082	10.1432	12.5774	14.3294	41.5482	0.263656	0.586824	0.105341	0.031190	0.012989	0.000000
38	lug	9.4055	8.2965	8.8225	9.4502	11.5951	18.7242	0.185132	0.322333	0.247791	0.181043	0.061946	0.001754
*39	lug	10.3779	11.3670	17.5119	27.2716	29.7038	42.3309	0.610363	0.372230	0.017237	0.000131	0.000039	0.000000
40	lug	2.4114	1.5039	3.4231	6.5431	11.5823	25.0973	0.301741	0.475000	0.181943	0.038234	0.003077	0.000004
*41	lug	2.4388	2.9245	3.1230	6.8816	6.4808	18.3115	0.365492	0.286697	0.259603	0.039642	0.048436	0.000131
42	lug	3.3488	2.6627	4.6745	6.5622	9.7969	25.0203	0.315958	0.445259	0.162838	0.063366	0.012573	0.000006
*43	lug	13.2896	14.6304	8.3795	11.4344	10.3806	22.5243	0.050051	0.025602	0.582951	0.126556	0.214345	0.000494
44	lug	3.7388	1.3831	4.5267	5.7940	12.5951	27.9355	0.188983	0.613684	0.127448	0.067629	0.002256	0.000001
45	lug	29.5412	25.0857	25.7076	27.4888	43.1629	52.6330	0.050328	0.466992	0.342188	0.140437	0.000055	0.000000
46	lug	6.7181	3.5494	7.9592	6.2778	14.4511	29.1790	0.130193	0.634829	0.069998	0.162253	0.002725	0.000002
*47	lug	3.3561	4.3754	8.9153	13.4205	18.6472	32.2363	0.598874	0.359763	0.037169	0.003907	0.000286	0.000000
48	lug	5.5090	2.7278	5.2162	5.1430	11.3570	32.1022	0.134598	0.540725	0.155818	0.161628	0.007231	0.000000
*49	lug	2.5265	4.3202	6.9181	14.4654	14.2151	32.6048	0.655928	0.267511	0.072984	0.001676	0.001900	0.000000
*50	lug	3.7695	3.9989	8.9132	11.6595	15.6946	28.7478	0.502532	0.448055	0.038389	0.009724	0.001293	0.000002
*51	lug	5.6232	7.3056	11.0554	18.0907	21.4494	28.4467	0.666821	0.287522	0.044098	0.001308	0.000244	0.000007
52	lug	12.7834	9.2735	15.1938	17.3571	29.3237	41.4331	0.139188	0.804932	0.041705	0.014139	0.000036	0.000000
53	lug	7.7379	4.7531	8.7110	8.2468	17.3996	36.2535	0.146070	0.649708	0.089798	0.113258	0.001166	0.000000
*54	lug	3.2526	5.1123	8.3867	17.1324	19.5002	31.8179	0.679057	0.267959	0.052124	0.000658	0.000201	0.000000
*55	lug	4.2480	6.9259	10.9053	19.7425	18.9774	32.8400	0.769809	0.201779	0.027591	0.000333	0.000487	0.000000
*56	lug	8.3051	5.0513	4.4275	4.9115	12.5009	30.5075	0.053711	0.273289	0.373321	0.293087	0.006591	0.000001
*57	lug	4.0068	4.3068	7.3374	7.9743	12.1603	20.4130	0.453593	0.390406	0.085790	0.062393	0.007694	0.000124
*58	lug	8.6496	5.5120	4.4303	7.5649	7.9470	25.8322	0.058182	0.279324	0.479735	0.100077	0.082671	0.000011
59	lug	7.5638	7.3943	13.8287	17.4020	22.1245	38.4315	0.467280	0.508605	0.020379	0.003414	0.000322	0.000000
60	lug	0.8736	3.6655	4.3791	6.1111	12.0297	33.7147	0.082634	0.456499	0.319502	0.134396	0.006969	0.000000

Plant position classification results *(cont from prev page)*

Data	PlPos	priming	lug	cutter	leaf	smoker	tip	priming	lug	cutter	leaf	smoker	tip
61	lug	11.2972	9.7095	11.9812	10.0673	11.4581	20.5227	0.149155	0.329914	0.105954	0.275868	0.137628	0.001480
62	lug	9.9352	7.5278	9.1795	7.6849	8.7821	22.5823	0.093861	0.312789	0.136956	0.289161	0.167066	0.000168
*63	lug	10.0479	7.8195	4.1087	4.2670	4.3897	26.0214	0.017105	0.052121	0.333269	0.307908	0.289590	0.000006
64	lug	4.6659	3.7427	5.9267	8.7047	13.3650	29.9395	0.306319	0.485990	0.163077	0.040658	0.003955	0.000001
*65	lug	12.5691	8.0896	7.0776	5.5903	11.0757	33.4927	0.016435	0.154347	0.256001	0.538537	0.034679	0.000000
66	lug	24.6985	20.0336	28.0046	33.8548	41.4481	60.3281	0.086917	0.895529	0.016641	0.000893	0.000020	0.000000
*67	lug	14.4749	13.9289	11.9721	20.1859	15.0217	36.9149	0.150888	0.198248	0.527391	0.008680	0.114792	0.000002
68	lug	9.4315	8.2633	15.3672	19.6717	29.5924	41.2512	0.350782	0.629073	0.018035	0.002096	0.000015	0.000000
69	lug	4.5858	3.7429	5.3575	6.3574	9.3543	21.8224	0.269637	0.410962	0.183314	0.111189	0.024849	0.000049
*70	lug	7.7553	7.5083	5.7355	7.0719	6.0974	13.0055	0.115639	0.130842	0.317473	0.162748	0.264923	0.008376
*71	cutter	6.6657	3.7094	2.3908	1.9279	6.4106	23.6385	0.038934	0.170719	0.330075	0.416033	0.044232	0.000008
72	cutter	6.9651	6.1773	3.3552	4.4866	7.4512	20.8359	0.078121	0.115833	0.474950	0.269754	0.061266	0.000076
*73	cutter	13.9467	13.0101	9.8685	11.5493	7.7483	12.8561	0.026665	0.042592	0.204881	0.088416	0.591445	0.046001
74	cutter	7.1790	6.1578	4.0237	7.5453	14.5397	28.1278	0.119510	0.199131	0.578836	0.099506	0.003013	0.000003
75	cutter	7.5157	7.6863	7.2781	10.8918	7.8853	30.0045	0.246275	0.226134	0.277338	0.045532	0.204718	0.000003
76	cutter	9.9437	6.0786	4.8492	6.2564	12.6030	29.4777	0.036679	0.253349	0.468470	0.231795	0.009704	0.000002
77	cutter	13.9937	11.3168	9.1842	11.9086	9.3958	30.5268	0.034856	0.132910	0.386053	0.098870	0.347302	0.000009
78	cutter	5.3924	3.9474	3.1621	4.6230	11.4780	24.7892	0.131123	0.270049	0.399924	0.192641	0.006255	0.000008
79	cutter	8.5419	8.5140	3.6539	8.7116	7.9782	23.6175	0.063381	0.064268	0.730082	0.058223	0.084012	0.000034
80	cutter	12.8346	14.8558	12.1137	21.3232	15.0519	35.0526	0.318235	0.115837	0.456340	0.004565	0.105018	0.000005
81	cutter	12.7165	11.4347	4.2199	6.7094	8.1338	23.4445	0.009715	0.018441	0.679909	0.195827	0.096063	0.000045
82	cutter	10.1637	8.1180	5.7877	11.2458	10.5437	25.9367	0.070884	0.197129	0.632080	0.041263	0.058617	0.000027
83	cutter	13.1312	10.7161	6.9657	8.0655	13.5796	36.5237	0.025283	0.084581	0.551625	0.318306	0.020205	0.000000
*84	cutter	32.7230	24.9470	18.0508	11.7863	23.2379	38.9384	0.000027	0.001323	0.041611	0.953927	0.003111	0.000001
85	cutter	9.5462	6.0960	3.0090	3.0212	11.2801	27.0841	0.016828	0.094462	0.442165	0.439470	0.007072	0.000003
86	cutter	6.9787	4.5573	1.6764	3.0907	8.2073	23.4430	0.038382	0.128804	0.543877	0.268162	0.020765	0.000010
*87	cutter	5.4804	7.1777	9.4684	12.6722	12.9736	30.5337	0.619125	0.264981	0.084295	0.016986	0.014610	0.000002
*88	cutter	9.4210	5.9072	6.4923	8.2093	15.3831	28.2247	0.076907	0.445625	0.332604	0.140954	0.003902	0.000006
89	cutter	23.1296	16.9123	15.3135	17.0809	26.3769	45.4561	0.010642	0.238279	0.529967	0.219014	0.002098	0.000000
*90	cutter	10.7241	7.2979	6.1806	4.6483	15.3730	26.2059	0.026881	0.149082	0.260643	0.560752	0.002630	0.000012
91	cutter	3.2057	2.6728	1.7585	5.4275	8.8240	27.4191	0.210229	0.274419	0.433461	0.069222	0.012668	0.000001
92	cutter	5.3587	4.7881	2.4826	8.8015	11.1122	27.9837	0.147545	0.196253	0.621508	0.026383	0.008309	0.000002
*93	cutter	8.5502	13.9128	16.4298	25.6932	26.5788	33.5572	0.918911	0.062924	0.017875	0.000174	0.000112	0.000003
*94	cutter	18.0449	17.5517	17.3565	13.5333	12.2554	7.8775	.005190	0.006642	0.007322	0.049527	0.093828	0.837491
95	cutter	5.4592	5.0022	3.3380	6.8320	8.1424	19.0358	0.169183	0.212612	0.488623	0.085163	0.044229	0.000191
96	cutter	6.2848	8.4014	5.1631	9.7934	6.7320	18.4724	0.245451	0.085183	0.430070	0.042470	0.196273	0.000554
97	cutter	7.8374	6.5076	5.6205	12.0118	10.5012	32.4119	0.157187	0.305617	0.476206	0.019496	0.041493	0.000001
98	cutter	14.9608	11.3481	5.3436	8.0796	15.7706	30.6207	0.006191	0.037693	0.758787	0.193196	0.004130	0.000002
99	cutter	13.4050	10.4731	5.8075	9.5685	15.3876	33.3450	0.017495	0.075785	0.781097	0.119129	0.006493	0.000001
100	cutter	6.7738	6.9892	2.9783	10.2935	9.8875	23.2410	0.111710	0.100305	0.745186	0.019222	0.023548	0.000030
101	cutter	12.9080	11.1946	9.6635	12.6937	12.9423	22.6445	0.095025	0.223821	0.481242	0.105771	0.093410	0.000730
*102	cutter	5.1823	7.6408	10.7661	17.0686	21.4486	35.5428	0.737071	0.215593	0.045186	0.001934	0.000216	0.000000
*103	cutter	8.1339	5.1734	2.8835	1.4696	5.4823	20.2878	0.019620	0.086214	0.270904	0.549341	0.073876	0.000045
104	cutter	7.3620	6.9362	5.9314	8.1004	6.3900	24.0349	0.151527	0.187482	0.309842	0.104749	0.246364	0.000036
105	cutter	12.3805	12.2831	7.1159	7.9594	7.2575	14.2812	0.026030	0.027328	0.361954	0.237406	0.337218	0.010063
*106	leaf	13.0493	10.0331	13.9128	11.0210	15.4769	34.0919	0.108441	0.489948	0.070419	0.298974	0.032215	0.000003
107	leaf	15.1520	11.3296	10.0254	6.3852	17.1566	28.4641	0.009880	0.066801	0.128224	0.791457	0.003626	0.000013
108	leaf	14.9022	13.5819	14.4173	7.8174	12.3563	23.2464	0.023616	0.045697	0.030095	0.815890	0.084338	0.000364
*109	leaf	4.8717	2.7598	3.9892	4.5149	9.6271	26.0304	0.148866	0.427943	0.231440	0.177938	0.013810	0.000004
110	leaf	18.4336	12.1454	10.6173	4.2845	14.1955	37.1414	0.000791	0.018356	0.039408	0.934860	0.006586	0.000000
111	leaf	11.8109	9.5851	11.0344	8.8234	12.6160	23.6287	0.093961	0.285949	0.138535	0.418475	0.062825	0.000255
112	leaf	18.1866	13.1625	8.6965	8.0674	21.1955	35.2788	0.003495	0.043099	0.402012	0.550616	0.000776	0.000001
*113	leaf	6.0220	3.5706	7.6680	11.7746	15.7049	30.0421	0.203671	0.693811	0.089434	0.011476	0.001608	0.000001
114	leaf	10.6420	9.1294	8.8075	6.9149	15.0601	24.9349	0.082032	0.174763	0.205286	0.528846	0.009008	0.000065
115	leaf	20.7944	14.0323	12.9009	12.6757	24.8370	42.9244	0.007131	0.209651	0.369141	0.413133	0.000945	0.000000
116	leaf	24.2124	19.7691	12.2580	7.5274	12.9934	18.6438	0.000204	0.001885	0.080601	0.858196	0.055804	0.003309
*117	leaf	10.8754	8.0191	10.1508	10.2904	12.8055	29.9737	0.120071	0.500812	0.172498	0.160868	0.045742	0.000009
118	leaf	19.2469	13.3158	10.3911	4.4080	11.4776	30.5299	0.000549	0.010657	0.045997	0.916078	0.026717	0.000002
119	leaf	15.6347	14.2103	11.9173	11.5640	22.7718	35.9483	0.058353	0.118953	0.374355	0.446692	0.001645	0.000002
120	leaf	32.5314	27.9202	20.5943	14.9493	21.1796	34.3324	0.000138	0.001380	0.053780	0.904512	0.040135	0.000056
121	leaf	24.2037	18.5926	14.6945	10.5129	17.7010	38.2852	0.000910	0.015046	0.105656	0.854889	0.023498	0.000001
122	leaf	13.5133	9.4273	6.3981	1.6121	6.7812	17.8667	0.002189	0.016884	0.076784	0.840496	0.063398	0.000248
123	leaf	22.2610	17.7803	11.6450	6.9636	10.7057	19.8262	0.000379	0.003564	0.076592	0.795680	0.122503	0.001281
124	leaf	19.7099	15.1346	9.3090	5.7897	7.9401	25.7536	0.000623	0.006136	0.112952	0.656309	0.223950	0.000030
125	leaf	21.6194	16.5338	11.2142	10.9559	23.0236	39.5112	0.002483	0.031570	0.451258	0.513458	0.001230	0.000000
126	leaf	18.0603	13.6873	9.2558	3.8208	7.8968	18.7200	0.000671	0.005978	0.054808	0.829932	0.108128	0.000483

Plant position classification results *(cont from prev page)*

Data	PlPos	priming	lug	cutter	leaf	smoker	tip	priming	lug	cutter	leaf	smoker	tip
*127	leaf	9.1678	4.7247	6.6603	5.6602	15.3768	29.8057	0.051160	0.471783	0.179233	0.295528	0.002294	0.000002
128	leaf	25.0565	22.2427	19.0421	14.5353	14.8236	28.1216	0.002598	0.010610	0.052564	0.500425	0.433241	0.000561
129	leaf	15.9269	14.7783	7.7061	5.1113	8.2866	17.9544	0.003004	0.005335	0.183170	0.670373	0.137027	0.001090
130	leaf	23.6204	21.4866	23.1959	15.1872	15.9190	31.8587	0.008334	0.024223	0.010305	0.565082	0.391921	0.000136
131	leaf	16.9781	13.7830	12.0623	6.0864	8.5648	30.7616	0.003159	0.015609	0.036897	0.732261	0.212071	0.000003
132	leaf	9.2652	6.7442	4.6243	4.0789	5.2440	19.1514	0.028126	0.099205	0.286335	0.376093	0.210040	0.000201
*133	leaf	15.4778	13.7177	11.6870	11.3400	3.7274	15.8356	0.002667	0.006431	0.017751	0.021115	0.949805	0.002230
*134	leaf	7.3151	4.7831	1.8304	2.6195	7.4458	27.2283	0.031777	0.112704	0.493283	0.332469	0.029767	0.000002
*135	leaf	17.0735	15.1335	7.1868	7.1102	4.8183	16.9173	0.001335	0.003522	0.187236	0.194544	0.611919	0.001444
136	leaf	17.5506	12.9270	9.2491	6.4604	11.4067	21.0837	0.002838	0.028648	0.180188	0.726575	0.061266	0.000485
137	leaf	13.2546	9.1697	5.4530	2.6333	4.1918	22.5664	0.002829	0.021806	0.139852	0.572742	0.262744	0.000027
*138	leaf	2.5579	2.7974	1.1420	4.1990	5.1811	19.4582	0.216135	0.191736	0.438720	0.095139	0.058223	0.000046
139	leaf	13.6852	10.7469	10.7497	5.1302	10.5279	23.4902	0.011548	0.050179	0.050111	0.832090	0.055987	0.000086
140	leaf	16.6350	12.2385	7.5959	3.8608	12.3279	26.4214	0.001419	0.012787	0.130286	0.843269	0.012228	0.000011
141	smoker	16.5730	13.3115	12.8600	8.1633	6.0564	19.2582	0.003677	0.018783	0.023539	0.246420	0.706620	0.000960
*142	smoker	16.6186	12.2446	9.2205	4.1315	6.5857	27.6261	0.001397	0.012444	0.056447	0.718946	0.210760	0.000006
143	smoker	21.9063	20.1836	11.4060	12.5529	8.4225	21.1555	0.000870	0.002058	0.165738	0.093405	0.736664	0.001266
*144	smoker	2.1208	3.9818	2.7432	8.2819	8.4905	21.9934	0.451615	0.178097	0.330831	0.020745	0.018690	0.000022
145	smoker	33.2114	33.7998	28.4709	35.7460	19.9349	32.9087	0.001286	0.000958	0.013759	0.000362	0.982139	0.001496
*146	smoker	12.6736	12.7091	6.0266	9.7801	7.8875	19.4514	0.022238	0.021846	0.617244	0.094492	0.243430	0.000750
147	smoker	21.5266	19.9084	19.2814	22.4767	10.4870	24.2535	0.003894	0.008746	0.011966	0.002422	0.971975	0.000996
148	smoker	16.4051	14.1949	11.5735	9.9339	7.0775	22.6582	0.006815	0.020579	0.076326	0.173263	0.722717	0.000299
149	smoker	24.7719	23.6191	17.8824	17.4247	9.3367	27.1059	0.000431	0.000767	0.013499	0.016970	0.968200	0.000134
150	smoker	17.7315	14.1021	9.9933	8.5320	4.8783	12.1109	0.001267	0.007779	0.060695	0.126029	0.783176	0.021053
151	smoker	24.1459	20.3901	16.7761	13.7411	11.8116	31.6166	0.001417	0.009264	0.056438	0.257393	0.675456	0.000034
*152	smoker	36.4983	31.7300	26.7560	17.5452	22.9253	35.1219	0.000071	0.000771	0.009267	0.926838	0.062912	0.000141
*153	smoker	17.0957	15.1834	8.0661	6.1334	6.7361	14.4899	0.001936	0.005038	0.176913	0.464975	0.344012	0.007126
*154	smoker	19.9950	16.4862	15.8162	8.6595	13.5354	30.9500	0.003035	0.017541	0.024522	0.878188	0.076702	0.000013
155	smoker	19.5143	14.2598	15.3394	14.9370	11.6714	28.7789	0.012012	0.166203	0.096876	0.118467	0.606325	0.000117
*156	smoker	39.7647	37.1248	32.4785	24.3067	26.0859	42.3018	0.000308	0.001152	0.011755	0.699378	0.287321	0.000087
*157	smoker	6.7819	8.2485	16.1395	19.4305	21.4435	34.6615	0.670207	0.321926	0.006226	0.001201	0.000439	0.000001
158	smoker	35.2752	41.2292	34.5415	35.7742	23.1646	23.9409	0.001391	0.000071	0.002008	0.001084	0.593121	0.402325
159	smoker	20.2519	21.3479	21.1972	18.6427	10.9090	18.3585	0.008782	0.005077	0.005474	0.019635	0.938398	0.022634
160	smoker	19.5183	17.1731	12.2727	10.1429	3.5002	10.2933	0.000307	0.000991	0.011489	0.033323	0.922981	0.030909
161	smoker	33.9993	34.7303	32.2833	30.9690	19.6380	23.1641	0.000646	0.000448	0.001524	0.002940	0.848847	0.145595
*162	smoker	37.5132	36.5165	32.1529	32.0910	16.6412	12.8833	0.000004	0.000006	0.000057	0.000059	0.132491	0.867384
163	smoker	34.6096	32.6891	21.9635	20.9847	18.4869	31.5670	0.000215	0.000563	0.119996	0.195755	0.682486	0.000986
164	smoker	20.3616	22.3229	12.7932	17.5786	8.0514	14.1128	0.001841	0.000690	0.080995	0.007401	0.867201	0.041872
165	smoker	42.0378	37.1401	38.0230	30.6814	23.3133	41.5009	0.000084	0.000968	0.000623	0.024462	0.973754	0.000109
166	smoker	34.1012	32.9381	22.9478	23.1736	10.9017	18.8363	0.000009	0.000016	0.002367	0.002114	0.977005	0.018490
167	smoker	23.1394	20.3251	15.2468	14.3930	11.3198	32.8160	0.001981	0.008092	0.102510	0.157100	0.730302	0.000016
168	smoker	22.0824	20.4103	15.6457	17.9470	10.3676	13.0761	0.002099	0.004844	0.052458	0.016600	0.734412	0.189587
169	smoker	17.4810	17.1808	9.9931	12.1056	5.6548	16.8687	0.002324	0.002700	0.098212	0.034154	0.859455	0.003156
*170	smoker	16.0350	15.1936	12.5705	8.8121	11.2139	18.1462	0.017642	0.026869	0.099733	0.653088	0.196529	0.006139
171	smoker	47.8890	45.0369	42.1283	40.4178	26.6356	33.8932	0.000024	0.000098	0.000420	0.000989	0.972647	0.025822
172	smoker	53.5655	52.9763	46.6559	53.1226	40.9853	47.9433	0.001692	0.002271	0.053552	0.002111	0.912241	0.028132
*173	smoker	11.0021	9.5940	7.4270	9.5358	12.1036	28.4898	0.085799	0.173476	0.512641	0.178604	0.049466	0.000014
*174	smoker	8.1028	10.4329	13.4620	18.2684	13.6500	29.6032	0.690054	0.215232	0.047333	0.004280	0.043087	0.000015
175	smoker	20.7580	23.3646	19.3008	27.0125	16.8701	18.3091	0.072589	0.019717	0.150415	0.003182	0.507126	0.246970
176	tip	43.0383	40.1331	33.4565	31.4205	18.7920	12.5896	0.000000	0.000001	0.000028	0.000078	0.043054	0.956839
177	tip	39.2828	41.2838	37.8654	35.2665	31.6730	8.5838	0.000000	0.000000	0.000000	0.000002	0.000010	0.999988
178	tip	27.0144	25.9871	20.4579	22.6135	16.6972	10.0720	0.000201	0.000335	0.005319	0.001810	0.034873	0.957462
179	tip	30.0648	33.8026	28.9714	31.7264	16.6060	15.3110	0.000410	0.000063	0.000708	0.000179	0.343095	0.655545
180	tip	26.8047	23.3013	19.5073	17.4100	8.1828	7.7175	0.000040	0.000229	0.001527	0.004357	0.439381	0.554466
181	tip	22.3543	24.7551	20.0384	18.8721	15.6015	7.3141	0.000531	0.000160	0.001690	0.003027	0.015532	0.979061
182	tip	32.9409	34.2394	30.5049	31.5344	16.7487	8.3413	0.000004	0.000002	0.000015	0.000009	0.014720	0.985249
183	tip	31.7967	34.2720	26.1335	28.5563	18.7919	9.3717	0.000013	0.000004	0.000227	0.000068	0.008921	0.990767
184	tip	34.5161	31.9087	26.5756	26.7902	19.5848	8.5316	0.000002	0.000001	0.000120	0.000108	0.003963	0.995798
185	tip	22.1388	21.6238	21.0913	19.1524	11.2673	8.2582	0.000787	0.001018	0.001328	0.003502	0.180539	0.812826
186	tip	63.5037	58.8767	54.9473	55.1602	41.7513	23.6830	0.000000	0.000000	0.000000	0.000000	0.000119	0.999880
*187	tip	28.3384	24.9811	25.5255	23.7371	20.6359	23.4360	0.012646	0.067765	0.051616	0.126221	0.595027	0.146725
188	tip	31.9921	33.7601	25.1112	23.3065	12.6912	12.2931	0.000029	0.000012	0.000902	0.002224	0.448973	0.547860
*189	tip	15.8344	17.7425	14.0334	18.4940	10.6973	14.6282	0.052671	0.020288	0.129619	0.013933	0.687217	0.096271
190	tip	67.5005	70.9013	64.7050	66.3518	50.3074	28.0051	0.000000	0.000000	0.000000	0.000000	0.000014	0.999986
191	tip	21.0840	22.7953	16.9991	17.9487	12.7939	4.7074	0.000272	0.000116	0.002097	0.001305	0.017173	0.979037
192	tip	60.0818	63.0176	57.2341	58.4655	49.5079	24.6447	0.000000	0.000000	0.000000	0.000000	0.000004	0.999996

Plant position classification results *(cont from prev page)*

Data	PlPos	priming	lug	cutter	leaf	smoker	tip	priming	lug	cutter	leaf	smoker	tip
193	tip	31.7664	31.0008	27.1799	23.9573	16.6792	15.0354	0.000160	0.000235	0.001586	0.007945	0.302324	0.687751
*194	tip	4.9154	5.8182	8.0417	8.2746	10.4770	18.1749	0.477102	0.303796	0.099943	0.088954	0.029575	0.000630
195	tip	70.1802	70.6423	71.4006	66.3379	58.0410	26.0395	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
196	tip	24.2185	24.8059	21.8260	19.3462	10.1729	4.8837	0.000059	0.000044	0.000195	0.000675	0.066257	0.932769
*197	tip	17.2500	15.4310	12.2701	12.7293	6.9135	14.0228	0.004860	0.012068	0.058616	0.046590	0.853465	0.024401
198	tip	130.4979	132.8037	131.8506	129.7732	111.4592	67.0776	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
199	tip	27.3417	32.3801	28.1716	27.4395	22.4561	13.3399	0.000899	0.000072	0.000594	0.000857	0.010348	0.987229
200	tip	39.2350	39.7377	33.5933	35.1260	29.1985	9.6949	0.000000	0.000000	0.000006	0.000003	0.000058	0.999932
201	tip	75.8714	77.2073	72.3923	71.3005	61.0561	22.0874	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
202	tip	37.4499	45.2312	36.2199	40.8740	31.9197	25.5403	0.002472	0.000051	0.004572	0.000446	0.039255	0.953204
203	tip	44.2676	43.2403	39.1501	36.8842	24.4186	6.8589	0.000000	0.000000	0.000000	0.000000	0.000154	0.999846
204	tip	34.2036	38.8036	31.3338	34.5861	17.5611	14.1089	0.000037	0.000004	0.000154	0.000030	0.151049	0.848726
205	tip	56.7523	55.2517	51.0134	48.4754	42.6603	30.0018	0.000002	0.000003	0.000027	0.000097	.001780	0.998091
206	tip	40.7738	39.1280	36.9300	30.5442	25.1022	16.7815	0.000006	0.000014	0.000041	0.001010	0.015345	0.983583
207	tip	35.9011	36.0761	30.2411	30.4250	23.1471	12.9975	0.000011	0.000010	0.000179	0.000163	0.006211	0.993427
208	tip	35.1806	37.3596	34.0135	34.1026	33.5562	15.3744	0.000050	0.000017	0.000090	0.000086	0.000113	0.999645
209	tip	40.4688	41.0251	39.9889	37.5598	29.6328	6.9800	0.000000	0.000000	0.000000	0.000000	0.000012	0.999988
210	tip	32.2489	31.2136	27.9128	27.5089	27.6106	14.4237	0.000134	0.000225	0.001172	0.001434	0.001363	0.995671

Appendix C

Image Data for Plant Position Classification

C.1 Images of primings

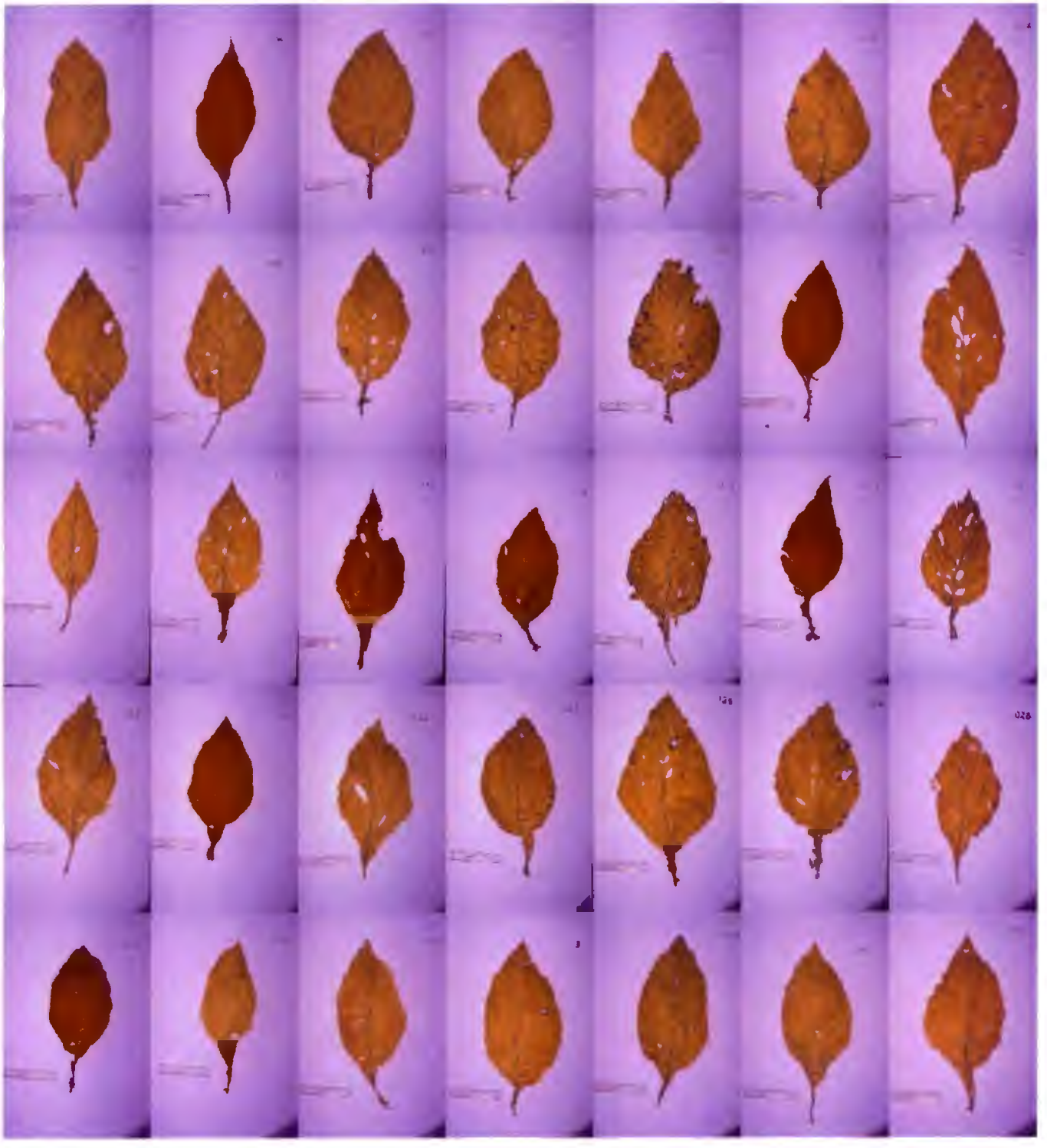


Figure C.1: **The 35 images of primings used in plant position classification**

C.2 Images of lugs

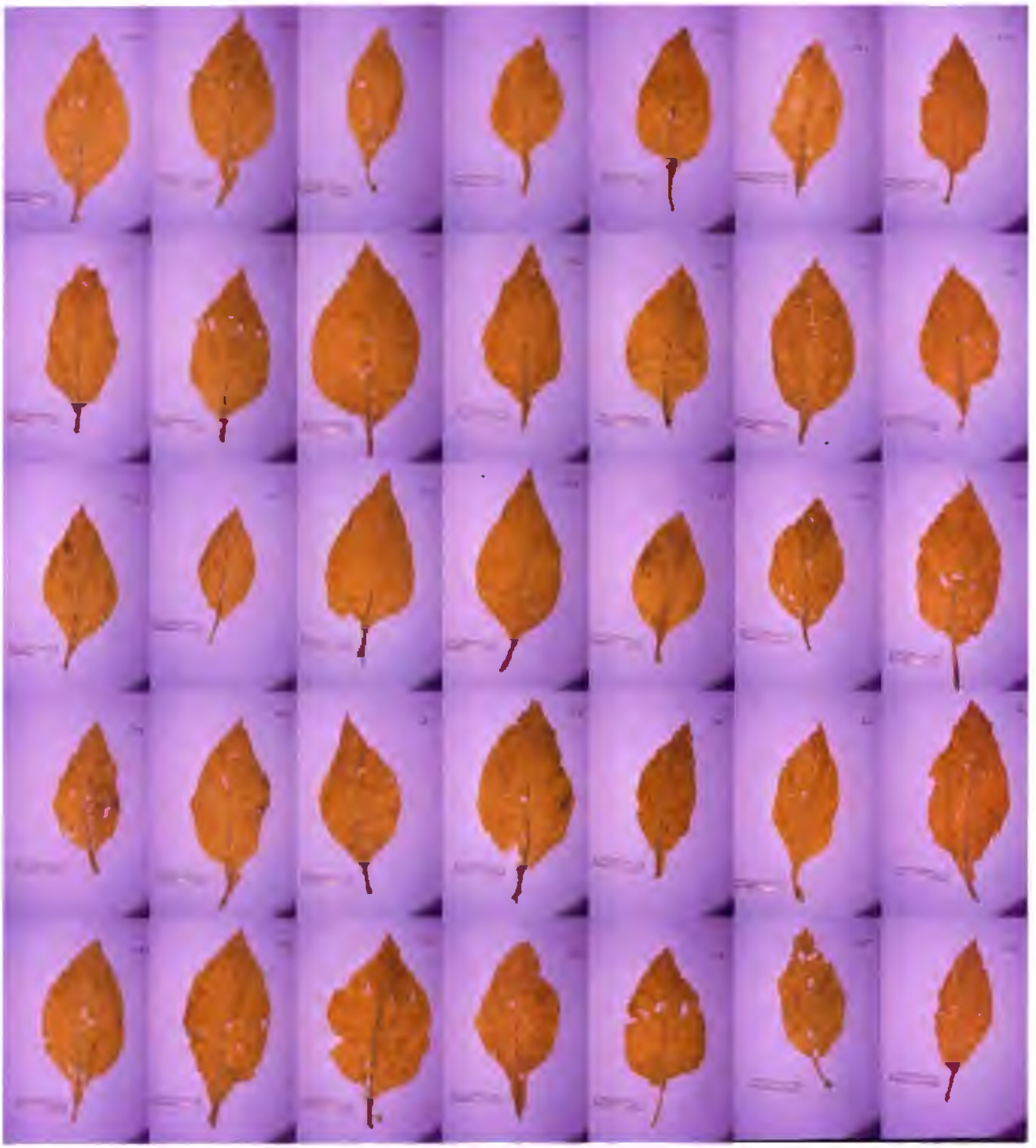


Figure C.2: **The 35 images of lugs used in plant position classification**

C.3 Images of cutters

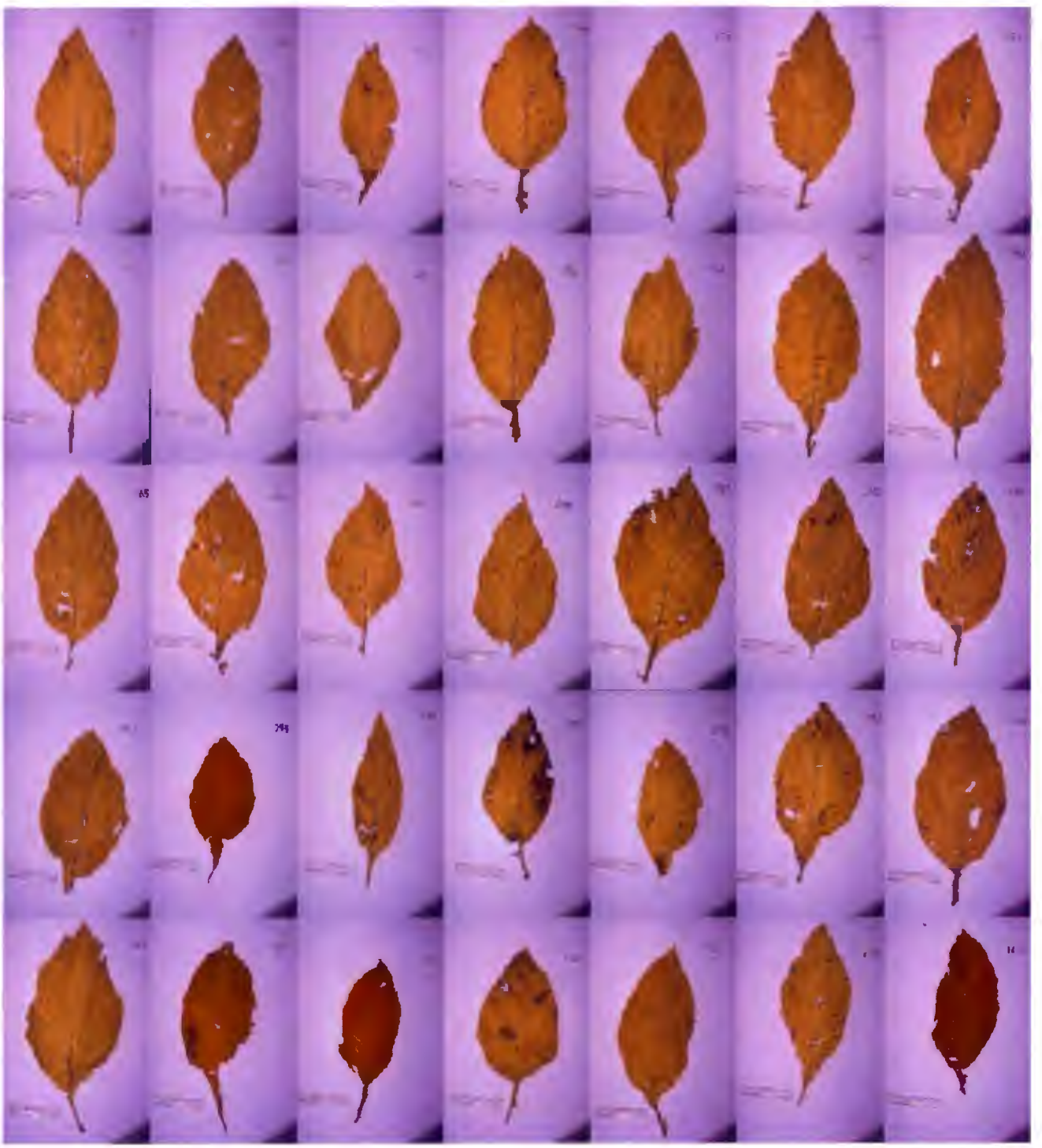


Figure C.3: **The 35 images of cutters used in plant position classification**

C.4 Images of leaf tobacco

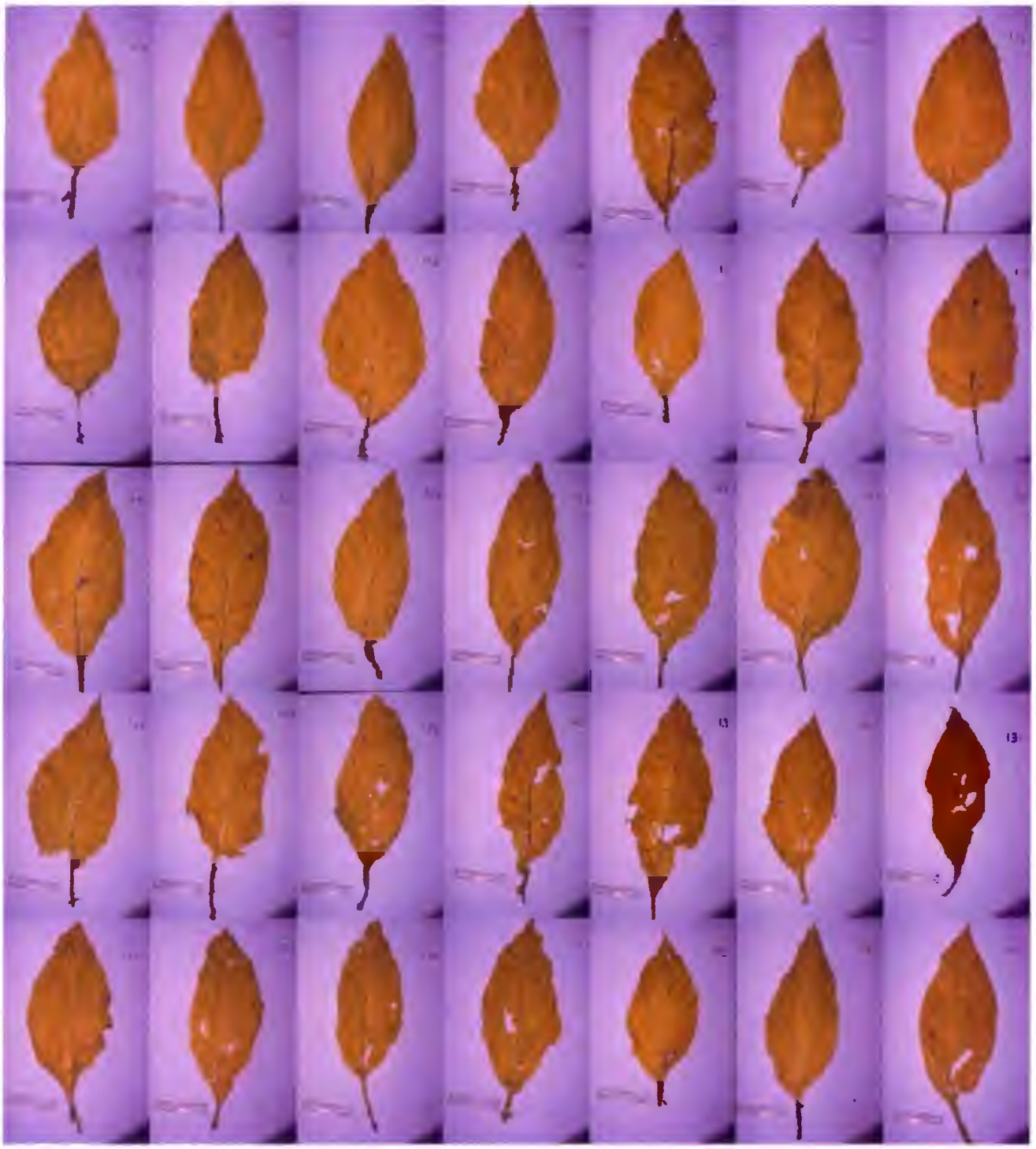


Figure C.4: **The 35 images of leaf used in plant position classification**

C.5 Images of smoking leaf

Figure C.5: **The 35 images of smoking leaf used in plant position classification**

C.6 Images of tips

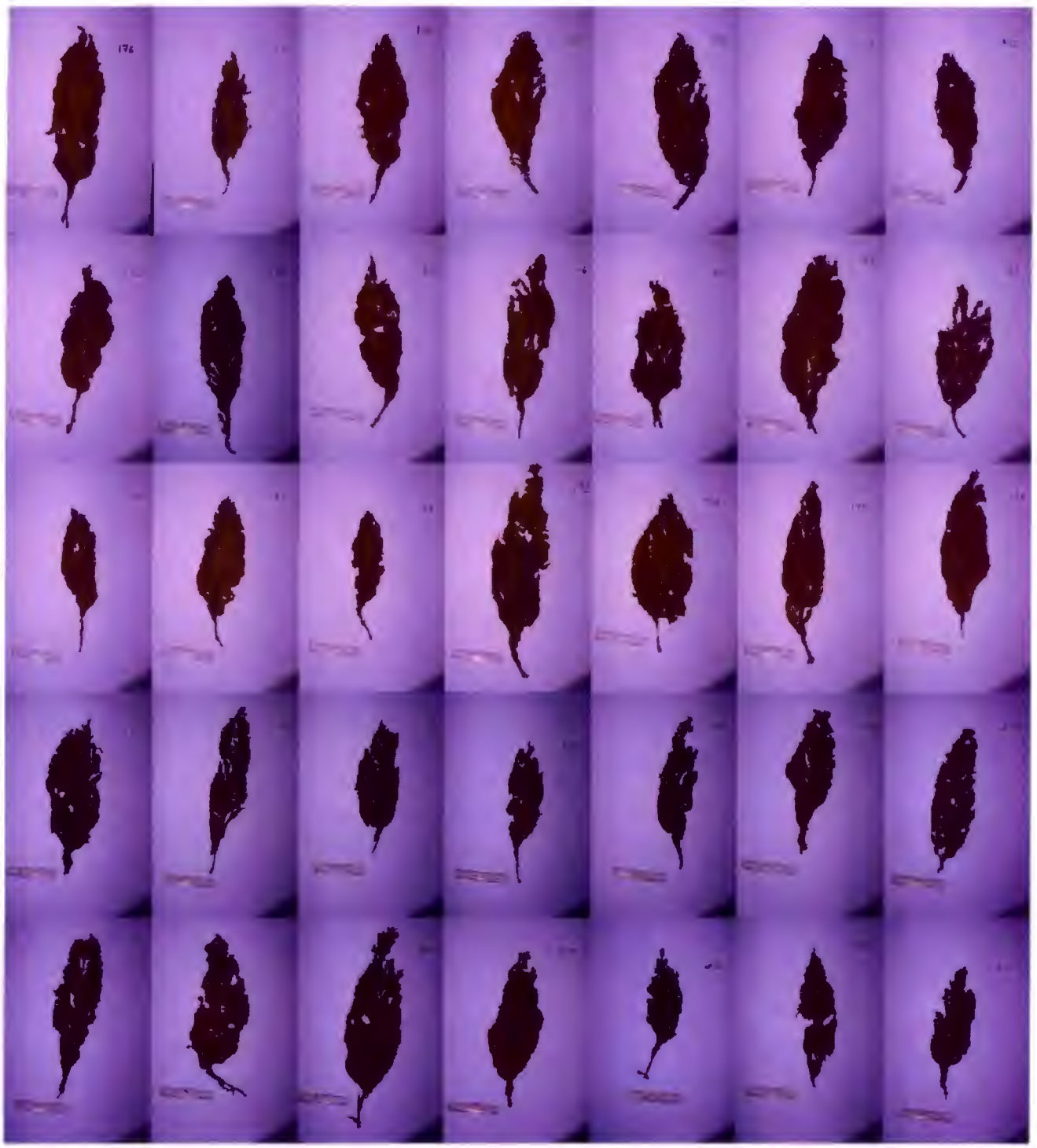


Figure C.6: **The 35 images of tips used in plant position classification**

Bibliography

[1] ABDALLAH, F.M. *Can Tobacco Quality be Measured?* Lockwood Publishing Co., Inc., 551 Fifth Avenue, New York, NY, USA, 1968, ch. 1. Can leaf quality be measured?, pp. 1–9.

[2] AITKEN, A.C. *Statistical Mathematics*, eighth ed. University Mathematical Texts. Oliver and Boyd, Ltd., Edinburgh, 1957.

[3] AKEHURST, B.C. *Tobacco*, second ed. Longmans, 1981.

[4] AKIMOTO, Y., WADA, Y., AND NIEDA, H. Studies on deciding the final harvest date of flue-cured tobacco cv. BY4 according to the colour scale #2. Bulletin 42, Okayama Tobacco Experiment Station, Tamashima, Okayama, Japan, 1983. pages 117-125.

[5] APPERSON, G.L. *The Social History of Smoking*. Martin Secker, London, 1914.

[6] ARIAS, E., ALONSO, A., AND BATTLE, J. Determination of leaf area in black tobacco variety 'corojo'. *Ciencia y Técnica en la Aricultura, Suelos y Agroquímica 6*, 2 (1982), 51–59. Instituto de Investigaciones de Suelos y Agroquímica.

[7] AUSTIN, G. A. *Perspectives on the History of Psychoactive Substance Use.* National Institute on Drug Abuse, 1979. See also *The Ibogaine Dossier* at http://www.ibogaine.desk.nl/drugmain.htm.

[8] BARCLAYS BANK. *Tobacco.* London: The Bank, Dominion, Colonial and Overseas, 1961.

[9] BISHOP, C. M. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[10] BRONGERS, G. A. *Nicotiana Tabacum.* H.J.W.Becht, Uitgeversmaat schappij N.V. Amsterdam, 1964.

[11] CAMPBELL, J. S. Trends in tobacco leaf usability. *Beiträge zur Tabakforschung International 16*, 4 (November 1995), 185–195. Presented at CORESTA: 15/9/93 and 4/10/93.

[12] CAROTENUTO, R. I colori del Virginia Bright (The colour of Virginia Bright tobacco). In *Informatori Agrario* (Scafati, Salerno, Italy, 1989), vol. 45 (35 supplemento) of *Sezione de Lavorazione e Transformazione*, Instituto Sperimentale per il Tabacco, pp. 34–38.

[13] CHAPMAN & HALL, Ed. *Dictionary of Organic Compounds*, 6 ed., vol. 5. Electronic Publishing Division, Chapman and Hall, 1996. Page 4738.

[14] COSTELLO, L. S. *Pilgrimage to Auvergne.* 1842.

[15] COUNT CORTI. *A History of Smoking.* George G. Harrap & Co., London, 1931. Translated by Paul England, page 50.

[16] COURT, W.A., BINNS, M.R. AND HENDEL, J.G. Chemical composition of representative grades of ontario flue-cured tobacco. *Canadian Journal of Plant Science 67*, 4 (1987), 1203–1219. Res. Sta., Agric. Canada, Delhi, Ontario.

[17] CRANDON, T. Progressive grower. *The Canadian Tobacco Grower 8*, 5 (July 1960), 10–11, 50–52.

[18] CRANDON, T. Vienna grower convinced good grading pays off. *The Canadian Tobacco Grower 9*, 9 (December 1961), 8–10.

[19] DANIELS, G.S. Grading and preparation of leaf for sale. *The Australian Tobacco Journal 2*, 1 (January 1961), 14–15, 21.

[20] DAVIES, E.R. *Machine Vision. Theory, Algortihms, Practicalities.*, second ed. Academic Press Limited, London, 1997.

[21] DE JONG, J. *Work Study Aspects of Flue-Cured Tobacco Production : Grading.* Farm Management and Work-Study Section, Department of Conservation and Extension, PO Box 8117, Causeway, Salisbury, Rhodesia, December 1978. Printed by the Government Printer, Salisbury (now Harare, Zimbabwe).

[22] DUDA, R. O., AND HART, P. E. *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., 1973.

[23] DUNHILL, A. H. *The Gentle Art of Smoking.* G.P.Putnam's Sons, New York, 1954.

[24] ECONOMIST. Book of Vital World Statistics. In *The Economist*, M. Smith-Morris, Ed. The Economist Books / Times Books / Random House, 1990. Page 239 : Drink and Tobacco.

[25] ERNST, A. On the etymology of the word tobacco. *The American Anthropologist 2* (April 1889), 133. Anthropological Society of Washington.

[26] EVENING NEWS. Newspaper article, 10th October 1914.

[27] FINANCIAL GAZETTE. Newspaper article, 10th September 1998.

[28] FOOD AND AGRICULTURE ORGANISATION OF THE UNITED NATIONS. Tobacco: supply, demand and trade projections, 1995 and 2000. *FAO economic and social development paper 86* (1990).

[29] FRALEIGH, J. B., AND BEAUREGARD, R. A. *Linear Algebra*, second ed. World student series edition. Addison-Wesley Publishing Company, 1990.

[30] GARNER, W. W. *The Production of Tobacco*, first revised ed. The Blakiston Company, 1951. Pages 186-193.

[31] GARVIN, R. T. It pays to maintain high standards on the floors. *Zimbabwe Tobacco Today*, 5 (May 1981), 17–19.

[32] GONZALEZ, R. C., AND WOODS, R. E. *Digital Image Processing.* Addison-Wesley Publishing Company, 1992.

[33] GOOCH, E.D. *Grading and pricing burley tobacco: the effect of light intensity and other factors.* PhD thesis, Agricultural Economics Department, University of Illinois, 1962.

[34] GOODSPEED, T.H. The genus nicotiana. *Chronica Botanica XVI* (1954).

[35] GRAHAM, S. *A Life of Peter the Great.* Ernest Benn, London, 1929.

[36] GRISE, V.N. The world tobacco market: government intervention and multilateral policy reform. Staff Report AGES 9014, USDA Economic Research Service, 1990.

[37] HERALD. In *The Herald* Newspaper, Harare, Zimbabwe, 9th September 1998.

[38] HOPE, K. *Methods of Multivariate Analysis*. University of London Press, 1968.

[39] JAHN, R., Ed. *Tobacco Dictionary*. Philosophical Library, New York, 1954.

[40] JÄHNE, B. *Digital Image Processing. Concepts, Algorithms and Scientific Applications*, second ed. Springer-Verlag, Berlin & Heidelberg, 1993.

[41] JAIN, R., KASTURI, R., AND SCHUNK, B. G. *Machine Vision*. McGraw-Hill series in Computer Science. McGraw-Hill, Inc., Singapore, 1995.

[42] KING JAMES I. *A Counter-Blaste to Tobacco*, 1954 ed. Rodale Books Inc., Emmaus, Pennsylvania and London, England, 1604.

[43] LEON-GARCIA, A. *Probability and Random Processes for Electrical Engineering*, second ed. Addison-Wesley, 1994.

[44] LINDEMAN, R. H., MARENDA, P. F., AND GOLD, R. Z. *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman and Co., 1980.

[45] LOW, A. *Introductory Computer Vision and Image Processing*. McGraw-Hill Publishing Company, Maidenhead, Berkshire, England, 1991.

[46] MACINNES, C. *The Early English Tobacco Trade*. Kegan Paul, Trench, Trubner & Co. Ltd, London, 1926.

[47] MBANGA, T. *Tobacco. A Century of Gold*. ZIL Publications (pvt) ltd., Harare, Zimbabwe, 1991.

[48] MCDONALD, E. Farm grading of flue cured tobacco. *The Australian Tobacco Journal 2*, 7 (July 1961), 9–12.

[49] MORRISON, N. *Introduction to Fourier Analysis*. John Wiley & Sons, Inc., 605, Third Avenue, New York, NY, USA, 1994.

[50] MUNSELL COLOR COMPANY INC. *Munsell Color Charts For Plant Tissues*, first ed. Baltimore, Maryland, USA, 1952.

[51] NAKAJIMA, T., AND KAKUBARI, T. *Estimation of Leaf Area in Burley Tobacco.* Bulletin 14, Morioka Tobacco Experiment Station, 1979. Pages 73-80.

[52] OVEIDO, GONZALO FERNÁNDEZ (Y VALDES). *Hystoria general y natural de las Indias*, vol. 1. Madrid Edition, 1851, Seville, 1535. Pages 130-131.

[53] OXFORD ENGLISH DICTIONARY. *The Compact Edition of the Oxford English Dictionary.* Clarendon Press, Oxford, 1971.

[54] RHODESIA TOBACCO ASSOCIATION. *Manual of Flue-Cured Virginia Tobacco Production Operations : Volume II.* Prepared by the Rhodesia Tobacco Association in conjunction with the Tobacco Research Board of Rhodesia and Nyasaland, PO Box 1302 , Bulawayo.

[55] RHODESIA TOBACCO ASSOCIATION. *R.T.A. Work Study Manual.*

[56] ROMER, I. F. *The Rhone, the Darro and the Guadaliquivir*, 1843.

[57] ROWSE, A. *Ralegh and the Throckmortons.* Macmillan and Co., London, 1962. p. 191.

[58] RUSHTON, W. *Vision as a Photic Process*, vol. II of *Photophysiology.* Academic Press, 1964, ch. 15, pp. 123–162. Editor : Arthur C. Giese.

[59] RUSS, J. C. *The Image Processing Handbook*, second ed. CRC Press, Inc. and IEEE Press, 1995.

[60] SALZMAN TOBACCO CO. *General Specification of Thailand Virginia Type Flue-cured Leaf Tobacco Grades.* Elia Salzman Tobacco Company Ltd in conjunction with the Thai Tobacco Leaf Development Co. Ltd, 1954.

[61] SHIGA, E., NAKAJIMA, T., AND KIMURA, S. *Investigations on the Development and the Practical Application of the Colour Plate for Estimating Leaf Colour of Air-Cured Tobacco.* Bulletin 17, Morioke Tobacco Experimental Station, 1983. pages 123-136.

[62] SOCIÉTÉ DE GENS DE LETTRES ET DE SAVANTS. *Bibliographie Universelle. Ancienne et Moderne.* L.G.Michaud, 1818.

[63] STAFF REPORTER. Cellar and Sorting Room. *Farming in S.A. 35*, 11 (November 1959), 33–35.

[64] STAFF REPORTER. Grading of Tobacco Improved with Artificial Lights. *The Bright Leaf 7*, 5 (Nov-Dec 1959), 16.

[65] STAFF REPORTER. Builds Grading Machine for $75. *The Canadian Tobacco Grower 8*, 8 (December 1960), 38–39.

[66] STAFF REPORTER. How Work Study Cuts Grading Costs. *Rhodesian Tobacco Journal 12*, 4 (April 1960), 41–3.

[67] STAFF REPORTER. Simple Colour Vision Test for Graders. *Rhodesian Tobacco Journal 13*, 11 (November 1961), 57.

[68] STAFF REPORTER. With a Commercial Leaf Grader. A System Developed to Simplify and Ensure Accurate Crop Grading. *Rhodesian Tobacco Journal 13*, 7 (July 1961), 59–61.

[69] STAFF REPORTER. Work Study Increases Grading Output. *Rhodesian Tobacco Journal 13*, 5 (May 1961), 55–56.

[70] STAFF REPORTER. Draycott's Grading System Increases Daily Output. *Rhodesian Tobacco Journal 14*, 5 (May 1962), 41–42.

[71] STAFF REPORTER. Fluorescent Lighting in Tobacco Grading Sheds. *Rhodesian Tobacco Journal 14*, 4 (April 1962), 41–42.

[72] STATSOFT, INC. *STATISTICA for Windows [Computer program manual]*. 2300 East 14th Street, Tulsa, OK 74104, USA, 1996. See also Statistica Help Files.

[73] TAIT, L. *The Leaf of the Petuns. Tobacco in Canada.* Ontario Flue-Cured Tobacco Growers' Marketing Board, Tillsonburg, Ontario, Canada, 1968. Page 117.

[74] TATTERSFIELD, G., AND FORBES, K. Classification of tobacco leaves by colour and plant position, by means of the computer processing of digital images. In *9th Annual Conference of the Pattern Recognition Association of South Africa* (1998).

[75] TILLEY, N. M. *The Bright Tobacco Industry 1860-1929.* University of Noth Carolina Press, 1948.

[76] TOBACCO INDUSTRY. *Tobacco Industry Profile 1997.* In Tobacco Resolution Online, 1997. TobaccoResolution.com was created by Philip Morris Incorporated, R.J. Reynolds Tobacco Company, Brown & Williamson Tobacco Corporation, Lorillard Tobacco Company and the United States Tobacco Company. It is published online at http://www.tobaccoresolution.com/0008.htm.

[77] TOBACCO INDUSTRY MARKETING BOARD. *Descriptions of Grades and Grade Symbols used in Classification of Zimbabwe Flue-Cured Tobacco.* Harare, Zimbabwe. Published by the Market Information Division of the TIMB.

[78] TOBACCO INDUSTRY MARKETING BOARD. *Statistical Summary of Flue-Cured Virginia Auction Sales from 1936 to Date.* Harare, Zimbabwe, 1996. Published by the Market Information Division of the TIMB.

[79] TOBACCO INDUSTRY MARKETING BOARD. *Weekly Flue-Cured Tobacco Marketing Report.* Harare, Zimbabwe, September 1997. Week 23 — Published by the Market Information Division of the TIMB.

[80] TOBACCO MARKETING BOARD OF ZIMBABWE. *The One Industry Approach to Grading and Presentation of Flue-Cured Tobacco.* Union Avenue, Harare, Zimbabwe, 1974. Revised and updated.

[81] TOBACCO MARKETING BOARD OF ZIMBABWE. *Annual Statistical Report — Zimbabwe Flue-Cured Tobacco Crop - 1994.* Harare, Zimbabwe, 1994. Published by the Market Information Division of the TMB.

[82] TOBACCO MARKETING BOARD OF ZIMBABWE. *The Tobacco Marketing Board definition of terms used in the tobacco leaf classification system (for Virginia flue-cured tobacco).* Harare, Zimbabwe, 1994. Published by the Market Information Division of the TMB.

[83] TOBACCO MARKETING BOARD OF ZIMBABWE. *Annual Statistical Report — Zimbabwe Flue-Cured Tobacco Crop - 1996.* Harare, Zimbabwe, 1996. Published by the Market Information Division of the TMB.

[84] TODD, FURNEY, A. *Flue-Cured Tobacco. Producing a Healthy Crop.* Parker Graphics, USA, 1981. pages 266-73.

[85] TUCKER, C., AND CHAKRABARTY, S. Quantitative assessment of lesion characteristics and disease severity using digital image processing. *Journal of Phytopathology 145*, 7 (July 1997), 273–278.

[86] UNITED STATES DEPARTMENT OF AGRICULTURE. *Grading, Standards and Labeling.* Standards Document E710. See http://waffle.nal.usda.gov/cc/d_e710.html.

[87] UNITED STATES DEPARTMENT OF AGRICULTURE. *Type Classification of American Grown Tobacco.* Miscellaneous Circular 55, United States Department of Agriculture, Washington, 1926. Superintendant of Documents.

[88] UNITED STATES DEPARTMENT OF AGRICULTURE. *Yearbook of Agriculture.* USDA, 1926. pages 716-7.

[89] UNITED STATES DEPARTMENT OF AGRICULTURE. *Yearbook of Agriculture.* USDA, 1954.

[90] UNITED STATES DEPARTMENT OF AGRICULTURE. *Official Standard Grades for Flue-Cured Tobacco U.S. Types 11,12,13,14 & Foreign Type 92*, March 1989.

[91] UNITED STATES DEPARTMENT OF AGRICULTURE. *Tobacco: World Markets and Trade.* Tech. Rep. FT-2-97, Foreign Agriculture Service, February 1997. http://www.fas.usda.gov/tobacco/circular/1997/9702/index.html.

[92] USA TODAY. *Highlights of tobacco settlement.* In USA Today Online, 1998. http://www.usatoday.com/news/smoke/smoke275.htm.

[93] USA TODAY. *Tobacco settlement is a done deal.* In USA Today Online, 1998. http://www.usatoday.com/news/smoke/smoke279.htm.

[94] VICTORIA DEPARTMENT OF AGRICULTURE. *Tobacco Growing in Victoria : Recommendations*, 1981. Page E-8.

[95] VOGES, E. *Tobacco Encyclopedia.* Tobacco Journal International, 1984. Page 149.

[96] WALDMAN, M. *Sir Walter Raleigh.* Collins, London, 1943.

[97] WEEKS, A. R. *Fundamentals of Electronic Image Processing.* SPIE/IEEE series on imaging science and engineering. SPIE Press and IEEE Press, 1996.

[98] WERNER, C. *A Textbook on Tobacco. An Exhaustive Technical Treatise on the Culture, the Manufacture and the Merchandising of Tobacco and Tobacco Products.* Tobacco Leaf Publishing Company, 1914. p. 86.

[99] WORDEN, B. *Stuart England.* Phaidon, Oxford, 1986.

[100] ZIMBABWE ONLINE. *Zimbabwe. The Week's News.* Online Statistics, September 1999. www.zimbabwe.8m.com/stats.html.

[101] ZIMBABWE TOBACCO ASSOCIATION. Tobacco trends 1/97. Supplement to *Zimbabwe Tobacco* (March 1997). Published by the Market Information Department of the ZTA.

[102] ZIMBABWE TOBACCO ASSOCIATION. Tobacco trends 2/97. Supplement to *Zimbabwe Tobacco* (June 1997). Published by the Market Information Department of the ZTA.

[103] ZIMBABWE TOBACCO ASSOCIATION. Tobacco trends 3/97. Supplement to *Zimbabwe Tobacco* (September 1997). Published by the Market Information Department of the ZTA.

[104] ZIMBABWE TOBACCO ASSOCIATION. Tobacco trends 4/98. Supplement to *Zimbabwe Tobacco* (December 1998). Published by the Market Information Department of the ZTA.